ABSTRACT

Title of dissertation: A THEORY OF CRAMÉR-RAO BOUNDS FOR CONSTRAINED PARAMETRIC MODELS

Terrence Joseph Moore, Jr.,
Doctor of Philosophy, 2010

Dissertation directed by: Professor Benjamin Kedem
Department of Mathematics

A simple expression for the Cramér-Rao bound (CRB) is presented for the scenario of estimating parameters $\boldsymbol{\theta}$ that are required to satisfy a differentiable constraint function $\boldsymbol{f}(\boldsymbol{\theta})$. A proof of this constrained CRB (CCRB) is provided using the implicit function theorem, and the encompassing theory of the CCRB is proven in a similar manner. This theory includes connecting the CCRB to notions of identifiability of constrained parameters; the linear model under a linear constraint; the constrained maximum likelihood problem, it's asymptotic properties and the method of scoring with constraints; and hypothesis testing. The value of the tools developed in this theory are then presented in the communications context for the convolutive mixture model and the calibrated array model.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Theory of Cramer-Rao Bounds for Constrained Parametric Models	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A simple expression for the Cram?er-Rao bound (CRB) is presented for the scenario of estimating parameters that are required to satisfy a differentiable constraint function f( ). A proof of this constrained CRB (CCRB) is provided using the implicit function theorem, and the encompassing theory of the CCRB is proven in a similar manner. This theory includes connecting the CCRB to notions of identifiability of constrained parameters; the linear model under a linear constraint the constrained maximum likelihood problem, it?s asymptotic properties and the method of scoring with constraints; and hypothesis testing. The value of the tools developed in this theory are then presented in the communications context for the convolutive mixture model and the calibrated array model.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**78**	

A THEORY OF CRAMÉR-RAO BOUNDS FOR CONSTRAINED PARAMETRIC MODELS

by

Terrence Joseph Moore, Jr.

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Radu Balan
Professor Paul Smith
Professor Doron Levy
Professor Prakash Narayan

© Copyright by
Terrence Joseph Moore, Jr.
2010

Acknowledgments

There are perhaps too many people to thank here without the regret of omitting someone important to me. So many have helped me become a mathematician. To the very talented high school math teachers at West Hills High School in Santee, CA, who encouraged the best students to strive further, to the many remarkable professors at American University who delighted in interacting with their students and no matter how difficult the material ensured the experience was always enjoyable, and to the excellent class of professors at the University of Maryland who challenged and pushed their students to excel. To all these people, I thank you.

I would like to acknowledge my undergraduate and Master's thesis advisor, Dr. Stephen Casey at American University, for his direction in the beginning of my college education. I do need to acknowledge the contribution of my first advisor, Dr. Dennis Healy, who encouraged the construction of this current work from the beginning. He allowed me the freedom to develop the theory using an approach I developed and generated a positive exuberance for whatever results I discovered. I only wish he could have survived to see its completion. I also need to acknowledge my second advisor, Dr. Benjamin Kedem, for his willingness to allow me to complete this document as it was originally conceived as well as for guiding me through the defense process and the selection of the committee.

I also need to acknowledge the financial support I have received from my employer, the U.S. Army Research Lab. Additionally, the professional support and mentoring I received from Dr. Brian Sadler and the advice from team members at

the lab has been invaluable in assisting me succeed through the process. I also need to thank Dr. John Gowens, Mr. Glenn Racine, Dr. Barbara Broome, Dr. Alex Kott, and Dr. Brian Rivera for allowing me the time at and away from work to complete this degree.

Table of Contents

List of Figures vii

List of Symbols and Abbreviations viii

1 INTRODUCTION 1
1.1 A note on the notation . . . 3

2 THE CRAMÉR-RAO BOUND 5
2.1 Definition . . . 6
2.1.1 Extensions . . . 7
2.2 Identifiability . . . 8
2.2.1 Local identifiability . . . 9
2.2.2 Strong Identifiability . . . 10
2.3 Linear Model . . . 11
2.3.1 Best Linear Unbiased Estimators . . . 12
2.3.2 Gaussian noise . . . 12
2.4 Maximum likelihood . . . 13
2.4.1 Efficient estimation . . . 14
2.4.2 Asymptotic Normality . . . 14
2.4.3 Scoring . . . 15
2.5 Hypothesis testing . . . 15
2.5.1 The Rao statistic . . . 15
2.5.2 The Wald statistic . . . 17
2.6 Discussion . . . 17

3 THE CONSTRAINED CRAMÉR-RAO BOUND 19
3.1 The Constrained CRB . . . 22
3.1.1 A proof of the CCRB . . . 25
3.1.2 Alternative formulas . . . 31
3.1.2.1 Gorman-Hero-Aitchison-Silvey CCRB . . . 32
3.1.2.2 Aitchison-Silvey-Crowder CCRB . . . 35
3.1.2.3 Xavier's & Barroso's Intrinsic Variance Lower Bound 36
3.1.3 Simple Extensions to the CCRB . . . 38
3.1.4 Properties of the CCRB . . . 41
3.1.5 Derivation of $\boldsymbol{g}(\boldsymbol{\xi})$ . . . 49
3.1.5.1 A Taylor series derivation . . . 50
3.1.5.2 A fixed point derivation . . . 51
3.2 Identifiability . . . 53
3.2.1 Local identifiability . . . 53
3.2.1.1 Local identifiability in the Aitchison-Silvey-Crowder CCRB formula . . . 57
3.2.2 Strong Identifiability . . . 58
3.3 Linear Model . . . 59

3.3.1 Best Linear Unbiased Estimation . . . 60
3.3.2 Uniform Minimum Variance Estimation under Gaussian noise 63
3.4 Constrained Maximum Likelihood Estimation . . . 66
3.4.1 Efficient estimation . . . 68
3.4.2 Asymptotic Normality . . . 69
3.4.3 The Method of Scoring Under Parametric Constraints . . . 71
3.4.3.1 Convergence Properties . . . 75
3.4.3.2 Linear constraints . . . 78
3.5 Hypothesis testing . . . 81
3.5.1 The Rao statistic . . . 82
3.5.2 The Wald statistic . . . 83
3.6 Discussion . . . 84

4 Applications of the CCRB in Communications Models 86
4.1 Convolutive Mixture Model . . . 88
4.1.1 Equivalent Convolutive Mixture Models . . . 90
4.1.1.1 The Vector-Matrix Model . . . 90
4.1.1.2 The Z transform model . . . 91
4.1.2 Strict Identifiability . . . 95
4.1.3 The Fisher information of the convolutive mixture model . . . 98
4.1.3.1 Complex-valued Fisher information . . . 98
4.1.3.2 CFIM for the convolutive mixture model . . . 99
4.1.3.3 Properties of the CFIM . . . 100
4.1.4 Constraints for the convolutive mixture model . . . 108
4.1.4.1 Pathways to regularity . . . 109
4.1.4.2 Norm channel + real-valued source constraint . . . 110
4.1.4.3 Semiblind constraints: $s^{(k)}(t) = p(t)$ for $t \in \mathbb{T}$ . . . 112
4.1.4.4 Unit Modulus constraint + Semiblind constraint . . . 113
4.2 Calibrated Array Model . . . 117
4.2.1 The Fisher information of the calibrated array model . . . 118
4.2.1.1 Indirect derivation of the FIM . . . 118
4.2.1.2 Direct derivation of the FIM . . . 120
4.2.1.3 Properties of the FIM . . . 121
4.2.2 Constraints for the calibrated array model . . . 122
4.2.2.1 Constraints on the complex-valued channel gain: $\mathbf{\Gamma} = \boldsymbol{I}_{K\times K}$ . . . 122
4.2.2.2 Semiblind constraints: $\boldsymbol{s}(t) = \boldsymbol{p}(t)$ for $t \in \mathbb{T}$ . . . 124
4.2.2.3 Finite alphabet constraint: $s^{(k)}(n) \in \mathbb{S}$ . . . 125
4.2.2.4 Unit modulus constraints: $|\boldsymbol{s}(n)| = 1$ for all n . . . 127
4.2.2.5 Unit modulus constraint; real-valued channel gain: $\mathrm{Im}(\gamma_k) = 0$ for all k . . . 129
4.2.2.6 Semi-blind and unit modulus constraint . . . 129
4.3 Discussion . . . 133

A Appendices 136
A.1 A proof of the CCRB using the Chapman-Robbins version of the Barankin bound 136
A.2 A proof of the CCRB using the method of implicit differentiation . . 138
A.3 Alternative proof of asymptotic normality 140

B Proofs of Convergence Properties of Constrained Scoring 142

C Proofs of Theorems in Chapter 4 147

Bibliography 161

List of Figures

3.1 Reparameterization of $\boldsymbol{f}(\boldsymbol{\theta}) = 0$ to $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$. 26

3.2 Projection of the observations $\boldsymbol{x}$ onto the linear space $\boldsymbol{H}\Theta$ and the linear constraint space $\boldsymbol{H}\Theta_f$. 61

3.3 Path created by iterates from the method of scoring with constraints. 74

4.1 Convolutive Mixture: Finite Impulse Response (tapped delay line) model. 89

4.2 Convolutive Mixture: Norm-constrained channel estimation performance. 111

4.3 Convolutive Mixture: Example of multipath channel. 115

4.4 Convolutive Mixture: Source estimation with varying $\Delta\psi$. 116

4.5 Calibrated array model geometries. 118

4.6 Calibrated Array: CCRBs on AOA for blind, constant modulus, and known signal models. 131

4.7 Calibrated Array: CCRBs on AOA for semiblind, constant modulus + semiblind, and known signal models. 132

4.8 Calibrated Array: CCRBs on signal phase for blind, constant modulus, semiblind, and constant modulus + semiblind. 134

List of Symbols and Abbreviations

Symbols

$\boldsymbol{b}(\boldsymbol{\theta})$	the bias of an estimator of $\boldsymbol{\theta}$
$\mathbb{C}$	the complex numbers
$\mathbb{C}^*$	$\mathbb{C} \bigcup \{\infty\}$
$\mathcal{CN}(\boldsymbol{a}, \boldsymbol{B})$	complex normal with mean $\boldsymbol{a}$ and (co)variance matrix $\boldsymbol{B}$
χ^2_r	chi-square distribution with r degrees of freedom
$\xrightarrow{d}$	convergence in distribution
$E_{\boldsymbol{\phi}}(\cdot)$	the expectation of $(\cdot)$ with respect to the density modeling $\boldsymbol{\phi}$
$\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$	the constraint function
$\boldsymbol{F}(\boldsymbol{\theta})$	the Jacobian (or gradient) of the constraint function at $\boldsymbol{\theta}$
$\boldsymbol{I}(\boldsymbol{\theta})$	the Fisher information for the parameter $\boldsymbol{\theta}$
$\boldsymbol{I}_{m\times m}$ or $\boldsymbol{I}_m$	the identity matrix of size m
$\mathrm{Im}(\cdot)$	the imaginary part of $(\cdot)$
$\boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta})$	the Fisher score
$\mathcal{N}(\boldsymbol{a}, \boldsymbol{B})$	normal with mean $\boldsymbol{a}$ and variance (covariance) matrix $\boldsymbol{B}$
$p(\boldsymbol{x}; \boldsymbol{\theta})$	the probability density function (pdf)
$\mathbb{R}$	the real numbers
$\mathrm{Re}(\cdot)$	the real part of $(\cdot)$
$\boldsymbol{t}(\boldsymbol{x})$	an estimator of some parameter (vector)
$\boldsymbol{\theta}$	the parameter
Θ	the parameter space
Θ_f	the constraint manifold (parameter space)
$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x})$	the ML estimator of $\boldsymbol{\theta}$ depending on the observation $\boldsymbol{x}$
$\boldsymbol{x}$	the observation (sample)
$(\cdot)_+$	$(\cdot)$ if the value is nonnegative, otherwise zero

Abbreviations

BER	bit-error rate
BLUE	best linear unbiased estimate (or estimator)
CCRB	constrained Cramér-Rao bound
CFIM	complex FIM
CLSE	constrained least squares estimate (or estimator)
CML	constrained maximum likelihood
CMLE	constrained maximum likelihood estimate (or estimator)
CRB	Cramér-Rao bound
FIM	Fisher information matrix
FIR	finite impulse response
iid	independent and identically distributed
LSE	least squares estimate (or estimator)
MIMO	multiple-input, multiple-output
ML	maximum likelihood
MLE	maximum likelihood estimate (or estimator)
MSE	mean-square error
MVUE	minimum variance unbiased estimate (or estimator)

NLB	nullity lower bound of the FIM (or CFIM)
SER	symbol-error rate
SIMO	single-input, multiple-output
SISO	single-input, single-output

Chapter 1

INTRODUCTION

The Cramér-Rao bound (CRB), the inverse of the Fisher information, is a limit on the performance of the estimation of parameters under certain conditions. Hence, the variance of any unbiased estimator cannot be lower than this bound, and the CRB may also be interpreted in some sense as a measure of performance potential. For a large number of scenarios, it is often of interest to gauge the performance of estimation under parametric constraints. The traditional approach in deriving CRBs for these cases is to find a reparameterization that represents the constraint. However, such an approach is not always feasible for a large class of constraints and numerical approximations are often restricted to the particular model. An alternative and equivalent option to reparameterization, the constrained Cramér-Rao bound, is presented herein, which is also computationally simple to program.

In communications design research, estimation performance metrics are often interpreted to represent performance potential to study the feasibility of a model to meet a certain measure of desired reliability. This approach is often practical since it avoids the necessity of searching for the best performer over a class of estimators for a particular trial model and since the CRB is analytically or numerically simple to compute. The downside to this approach is that for each model, the CRB needs

to be derived each time. This effectively prohibits the practitioner from studying an overly large class of models or, in the context of this work, a large class of constraints on some base model. This restriction is to a great extent eliminated using the constrained Cramér-Rao bound, as the Fisher information, which involves an integration over many variables, for the base model only needs to be evaluated once.

Chapter 2 offers a quick review of several connections with the Cramér-Rao bound (CRB) within mathematical statistics and serves as a reference point for the study of parameters under parametric equality constraints, discussed in Chapter 3. With the possible exception of the identifiability relationship in section 2.2, much of this section is familiar and well represented in standard mathematical statistics texts.

In Chapter 3, a general theory of the constrained Cramér-Rao bound (CCRB) is presented. In section 3.1, the CCRB is defined and proven, alternative formulas are presented, and several interesting properties of the bound are detailed. In section 3.2, a connection is made between the CCRB and two different notions of identifiability under certain conditions. In section 3.3, the linear model with linear constraints is examined in the context of the CCRB. In section 3.4, connections between the CCRB and constrained maximum likelihood estimation are detailed, including an asymptotic normality result and an adaptation of the method of scoring to the constrained parameter scenario. This chapter concludes with the consideration of hypothesis testing under constraints in section 3.5. Chapters 2 and 3 are designed so that the section numbers correspond directly, i.e., section 3.x relates

a concept for constrained parameters that section 2.x reviews for unconstrained parameters.

In Chapter 4, the analytic tools developed in Chapter 3 are applied in the communications context of the convolutive mixture model (section 4.1) and the calibrated array model (section 4.2). These models are defined and their Fisher information matrices developed in section 4.1.3 and section 4.2.1, respectively. A variety of constraints for these models are considered in sections 4.1.4 and 4.2.2.

1.1 A note on the notation

All elements will be denoted in lowercase math font: a. All vectors will be column vectors and be denoted in a lowercase bold math font: $\boldsymbol{a}$. Hence, the ith element of the column vector $\boldsymbol{a}$ will be denoted as a_i. (This should not be confused with $\boldsymbol{a}_i$, which is often used as a subvector of the vector $\boldsymbol{a}$ and will be defined in context.) All matrices will be denoted in an uppercase bold math font: $\boldsymbol{A}$. All scalars, vectors, and matrices are assumed to have elements with real-valued numbers unless otherwise noted as complex-valued (where the complex number $i = j = \sqrt{-1}$ should be clear from context).

For vectors and matrices, $(\cdot)^T$ will denote the transpose operator (do not with confuse $(\cdot)'$, which is occasionally used here as a dummy variable), $(\cdot)^*$ will denote the conjugate operator (do not confuse with $(\cdot)^\star$, which is occasionally used as a variant of another vector or matrix), and $(\cdot)^H$ will denote the Hermitian (or conjugate transpose) operator. When a vector depends on another vector value as in $\boldsymbol{a}(\boldsymbol{\theta})$, then

the Jacobian will be a matrix denoted as $\boldsymbol{A}(\boldsymbol{\theta})$ where the ith row is the transposed of the vector $\frac{\partial}{\partial \boldsymbol{\theta}} a_i(\boldsymbol{\theta})$ and $a_i(\boldsymbol{\theta})$ is the ith row element of $\boldsymbol{a}(\boldsymbol{\theta})$. For square matrices, $(\cdot)^{-1}$ will denote the inverse of the matrix and $(\cdot)^{\dagger}$ will denote the pseudoinverse of the matrix. Of course, $(\cdot)^2$ will denote the square of the matrix.

For symmetric matrices (or Hermitian matrices in the complex-valued case, the expression $\boldsymbol{A} > \boldsymbol{B}$ will denote that the matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. Similarly, $\boldsymbol{A} \geq \boldsymbol{B}$ will denote that the matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.

Sets will be denoted in an uppercase blackboard math font: $\mathbb{A}$ or in an uppercase Greek letter math font: Θ. Also, all sets will be assumed open sets unless otherwise noted.

For convenience to allow the reader to find referenced items, numbered theorems, corollaries, and examples share numbering. Thus, the theorem immediately following Example x.4 in Chapter x is numbered Theorem x.5 even though the previous theorem is Theorem x.1.

Chapter 2

THE CRAMÉR-RAO BOUND

The Cramér-Rao bound (CRB) is a lower bound on the error covariance of any unbiased estimator under certain regularity conditions. As such, it is a measure of the optimal performance of an unbiased estimator under a given model. There are other mean-square error lower bounds, such as the Ziv-Zakai bound [78], the Hammersley-Chapman Robbins bound [26, 16], the Barankin bound [8], the Bhattacharyya bound [11], etc., and indeed, depending on the model, there are numerous other possible performance measures, such as classification bounds like bit-error rate (BER) or symbol-error rate (SER), but the CRB remains a very popular benchmark due to it's computational simplicity and its underlying well-developed theory.

This theory has led to numerous connections in areas of mathematical statistics, e.g., identifiability, linear models, maximum likelihood, including asymptotic normality and the method of scoring, and hypothesis testing. This short chapter is a quick review of just a few of these connections. A more complete discussion of the utility and application of the topics discussed in this chapter, as well as proofs of the definitions, theorems and statements herein, may be found in many standard statistical inference texts, such as Shao [62] or Casella and Berger [14] for statisticians, Kay [36] for signal processors, or Van Trees [72] for engineers.

2.1 Definition

Suppose we have an observation $\boldsymbol{x}$ in $\mathbb{X} \subset \mathbb{R}^n$ from a probability density function (pdf) $p(\boldsymbol{x};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is in an open set $\Theta \subset \mathbb{R}^m$ is a vector of deterministic parameters. The Fisher information matrix (FIM) of this model is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) \triangleq E_{\boldsymbol{\theta}}\left\{\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{s}^T(\boldsymbol{x};\boldsymbol{\theta})\right\}$$

where $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ is the Fisher score defined by

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}) \triangleq \left.\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$$

and the expectation is evaluated at $\boldsymbol{\theta}$, i.e., $E_{\boldsymbol{\theta}}(\cdot) = \int_{\mathbb{X}} (\cdot) p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}$. And suppose as *regularity conditions*, the pdf is differentiable with respect to $\boldsymbol{\theta}$ and satisfies

$$\frac{\partial}{\partial \boldsymbol{\theta}^T} E_{\boldsymbol{\theta}}(\boldsymbol{h}(\boldsymbol{x})) = E_{\boldsymbol{\theta}}(\boldsymbol{h}(\boldsymbol{x})\boldsymbol{s}^T(\boldsymbol{x};\boldsymbol{\theta})) \tag{2.1}$$

for $\boldsymbol{h}(\boldsymbol{x}) \equiv 1$ and $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{t}(\boldsymbol{x})$ where $\boldsymbol{t}(x)$ is an unbiased estimator of $\boldsymbol{\theta}$ [62]. These conditions are assured under a number of scenarios, for example, when the Jacobian and Hessian of the density function $p(\boldsymbol{x};\boldsymbol{\theta})$ is absolutely integrable with respect to both $\boldsymbol{x}$ and $\boldsymbol{\theta}$ [72], and essentially permit switching between the order of integration and differentiation. Under these assumptions, we have the following information inequality theorem [14, 62, 36, 72], independently developed by Cramér [17] and Rao [56].

Theorem 2.1. The Cramér-Rao bound is the inverse of the FIM,

$$\mathrm{CRB}(\boldsymbol{\theta}) \triangleq \boldsymbol{I}^{-1}(\boldsymbol{\theta}), \tag{2.2}$$

if it exists, and the variance of any unbiased estimator $\boldsymbol{t}(\boldsymbol{x})$ satisfies the inequality

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \mathrm{CRB}(\boldsymbol{\theta})$$

with equality if and only if $\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} = \mathrm{CRB}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ in the mean-square sense.

Example 2.2. Let $\boldsymbol{x} \sim \mathcal{CN}(\vartheta, \sigma^2)$ with unknown complex-valued mean ϑ and known variance σ^2. In terms of real-valued parameters, the equivalent model is

$$\begin{bmatrix}\mathrm{Re}(\boldsymbol{x})\\ \mathrm{Im}(\boldsymbol{x})\end{bmatrix} \sim \mathcal{N}(\begin{bmatrix}\mathrm{Re}(\vartheta)\\ \mathrm{Im}(\vartheta)\end{bmatrix}, \frac{\sigma^2}{2}\boldsymbol{I}_{2\times 2}).$$

From a well-known result on normal distributions [36, equation (3.31)],

$$\boldsymbol{I}(\begin{bmatrix}\mathrm{Re}(\vartheta)\\ \mathrm{Im}(\vartheta)\end{bmatrix}) = \frac{2}{\sigma^2}\boldsymbol{I}_{2\times 2}$$

and the CRB is $\frac{\sigma^2}{2}\boldsymbol{I}_{2\times 2}$.

2.1.1 Extensions

The performance of a function of parameters, e.g. the transformation $\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta})$, is often of more interest than the performance of the parameters. If the Jacobian of the transformation function is $\boldsymbol{K}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{k}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$, then the CRB on the performance of an unbiased estimator of α is [62, 36]

$$\mathrm{CRB}(\boldsymbol{\alpha}) = \boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{K}^T(\boldsymbol{\theta}),$$

i.e., if $\boldsymbol{S}(\boldsymbol{x})$ is an unbiased estimator of $\boldsymbol{\alpha}$, then $\mathrm{Var}(\boldsymbol{S}(\boldsymbol{x})) \geq \boldsymbol{K}(\boldsymbol{\theta})\mathrm{CRB}(\boldsymbol{\theta})\boldsymbol{K}^T(\boldsymbol{\theta})$. Implicit in this inequality for the transformation is that $\boldsymbol{\alpha}$ is differentiable with respect to $\boldsymbol{\theta}$ and (2.1) must also be satisfied for $\boldsymbol{h}(x) \equiv \boldsymbol{S}(\boldsymbol{x})$. Consequently, if

an estimator $\boldsymbol{t}(\boldsymbol{x})$ is biased with bias $\boldsymbol{b}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta}$, then the transformation formula above can be used to attain a bound for $\boldsymbol{\alpha} = \boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})$. Then $\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \mathrm{CRB}(\boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta}))$ where

$$\mathrm{CRB}(\boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})) = (\boldsymbol{I}_m + \boldsymbol{B}(\boldsymbol{\theta}))\,\mathrm{CRB}(\boldsymbol{\theta})\left(\boldsymbol{I}_m + \boldsymbol{B}^T(\boldsymbol{\theta})\right)$$

with $\boldsymbol{B}(\boldsymbol{\theta}) = \left.\frac{\partial \boldsymbol{b}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}\right|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$.

Theorem 2.1 requires a nonsingular Fisher information, however, there are a number of interesting cases where this requirement can not be met yet the model is still of interest. For this scenario, the pseudoinverse of the FIM is occasionally used as a bound in place of the CRB, i.e.,

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \boldsymbol{I}^{\dagger}(\boldsymbol{\theta})$$

for an unbiased estimator $\boldsymbol{t}(\boldsymbol{x})$. This bound inequality is trivial for nonidentifiable functions of the parameters [66], i.e., the variance is finite only if $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{I}^{\dagger}(\boldsymbol{\theta})$ where $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{K}(\boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{\theta})$ and $\boldsymbol{t}(\boldsymbol{x})$ is a biased estimator of $\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta})$ with bias $\boldsymbol{b}(\boldsymbol{\theta})$.

2.2 Identifiability

The ability to identify parameters determines the validity and utility of certain structural models. Criteria on the identifiability of parameters has numerous connections to parametric statistical measures, such as Kullback-Leibler distance [13] and the Fisher information matrix [58, 29, 69]. In this section, two of these connections are developed to establish conditions under which a particular parametric model is identifiable.

2.2.1 Local identifiability

To proceed in examining the identifiability criterion from the CRB, a definition of identifiability is required. A parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ is *identifiable* in the model $p(\boldsymbol{x}; \cdot)$ if there is no other other $\boldsymbol{\theta}' \in \Theta$ such that $p(\boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}; \boldsymbol{\theta}')$ for all $\boldsymbol{x} \in \mathbb{R}^n$. A parameter is *locally identifiable* if there exists an open neighborhood of $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}$ is identifiable in that neighborhood.

A parameter is (locally) identifiable in the additive noiseless case if the parameter is solvable (locally). *Estimable* parameters, i.e., expected values of functions of the observations [57], are also identifiable.[1] Hence, a non-identifiable parameter is not estimable regardless of the scheme or the number of observations. These scenarios exist when some inherent ambiguity exists in the model.

Example 2.3. The parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ in the model

$$x = \theta_1 + \theta_2 + w,$$

where w represents the observation noise, is not identifiable since it is indistinguishable from the parameter vector $\boldsymbol{\theta}' = [\theta_1 + a, \theta_2 - a]^T$ for any real number a.

The Fisher information matrix is called *regular* if $\boldsymbol{I}(\boldsymbol{\theta})$ is full rank, and $\boldsymbol{\theta}$ is said to be a *regular point* of $\boldsymbol{I}(\boldsymbol{\theta})$. If the FIM is singular, $\boldsymbol{\theta}$ is a *singular point*.

Example 2.4. The FIM for the model

$$x = \theta^2 + w,$$

[1] The converse is not true. While it is possible to develop estimation schemes for any identifiable parameters, there is no guarantee that those estimators will be unbiased.

where $w \sim \mathcal{N}(0,1)$, is $I(\theta) = 4\theta^2$. For any $\theta \neq 0$, θ is locally identifiable but not identifiable and at the same time a regular point of the FIM. For $\theta = 0$, θ is a singular point, yet is identifiable.

Fisher information-regularity implies local identifiability, but as the example demonstrates the converse is not true. Rothenberg [58] found a connection between local identifiability and the FIM under certain conditions.

Theorem 2.5 (Rothenberg)**.** Assume the FIM $\boldsymbol{I}(\boldsymbol{\theta})$ has constant rank locally about $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}$ is locally identifiable if and only if $\boldsymbol{I}(\boldsymbol{\theta})$ is regular.

2.2.2 Strong Identifiability

Suppose that $\boldsymbol{p}(\boldsymbol{x};\boldsymbol{\theta})$ is a normal pdf with mean $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathbb{R}^p$ and variance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, whose elements may be explicitly defined by a map $\boldsymbol{\varphi} : \Theta \to \mathbb{R}^q$ where $q \leq p + p(p+1)/2$ and it is assumed $m \leq q$. Then by the given definitions, (local) identifiability holds when $\boldsymbol{\varphi}$ is injective (locally) and since by a transformation on the FIM [12, p. 157]

$$\boldsymbol{I}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\varphi}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{I}(\boldsymbol{\varphi}(\boldsymbol{\theta})) \frac{\partial \boldsymbol{\varphi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$$

regularity holds when the Jacobian $\frac{\partial \boldsymbol{\varphi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ has full rank m.

Suppose there exists a set of indices $i_1, \ldots, i_m \in \{1, \ldots, q\}$ that make $\boldsymbol{\varphi}^*(\boldsymbol{\theta}) = [\boldsymbol{\varphi}_{i_1}(\boldsymbol{\theta}), \ldots, \boldsymbol{\varphi}_{i_m}(\boldsymbol{\theta})]^T$ injective on Θ. Then each $\boldsymbol{\theta} \in \Theta$ is *strongly identifiable* and $\boldsymbol{\varphi}^*$ is a *representative mapping*. By this definition, if $\boldsymbol{I}(\boldsymbol{\theta})$ is regular at $\boldsymbol{\theta}$ then $\boldsymbol{\theta}$ is in a strongly identifiable open neighborhood, and if $\boldsymbol{\theta}$ is strongly identifiable on Θ then it is also identifiable on Θ. The converses are not generally true. The following

theorem establishes conditions under which the converses are true [29].

Theorem 2.6 (Hochwald and Nehorai)**.** Let $\boldsymbol{\varphi} : \Omega \to \mathbb{C}^q$ be a holomorphic mapping of $\boldsymbol{z} \in \Omega \subset \cup_{\alpha \in A} \Omega_\alpha$, where $\Theta \subset \Omega \subset \mathbb{C}^m$ and Ω_α is open in $\mathbb{C}^m$ for each α. Then

(a) if $\boldsymbol{I}(\boldsymbol{z})$ is regular, there exists a strongly identifiable open neighborhood about $\boldsymbol{z}$, and

(b) if there exists a representative mapping $\boldsymbol{\varphi}_\alpha^* : \Omega_\alpha \to \mathbb{C}^q$ for each α, $\boldsymbol{I}(\boldsymbol{z})$ is regular for every $\boldsymbol{z} \in \Omega$.

Therefore, the existence of a proper holomorphic function(s) equates Fisher information regularity with strong identifiability for normal distributions. And locally constant rank in the FIM equates regularity with local identifiability for arbitrary distributions.

2.3 Linear Model

Suppose we have observations $\boldsymbol{x}$ from a linear model

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}, \tag{2.3}$$

on $\boldsymbol{\theta}$, where $\boldsymbol{H}$ is an observation matrix consisting of known elements and $\boldsymbol{w}$ is the noise from the observations with mean zero and variance $\boldsymbol{C}$.

2.3.1 Best Linear Unbiased Estimators

The Gauss-Markov theorem [14] states that the best linear unbiased estimator (BLUE) is given by the (weighted) least squares solution

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x}, \tag{2.4}$$

so called for minimizing the (weighted) least squares $(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})$. For any other LUE of $\boldsymbol{\theta}$, i.e. $\boldsymbol{A}\boldsymbol{x}$, then $\mathrm{Var}(\boldsymbol{A}\boldsymbol{x}) \geq \mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}))$ with equality if and only if $\boldsymbol{A} = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}$. This assumes a full column rank observation matrix $\boldsymbol{H}$. Otherwise, for any estimable function of the parameters $\boldsymbol{d}^T\boldsymbol{\theta}$, where $\boldsymbol{d}$ is in the column space of the transposed observation matrix $\boldsymbol{H}^T$, its least squares solution is $\boldsymbol{d}^T\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x})$ where

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{\dagger}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x}, \tag{2.5}$$

and $(\cdot)^{\dagger}$ is the generalized pseudoinverse of $(\cdot)$ [57, theorems 11.2B,11.3A-D]. This solution is also the BLUE with variance $\boldsymbol{d}^T\left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{\dagger}\boldsymbol{d}$.

2.3.2 Gaussian noise

Additionally, if the noise has a Gaussian distribution, i.e., $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, then the least squares solution is also the maximum likelihood estimator (MLE) and the minimum variance unbiased estimate (MVUE).

Theorem 2.7. If the observations obey the linear model in (2.3), where $\boldsymbol{H}$ is a known full column rank matrix, $\boldsymbol{\theta}$ is an unknown parameter vector, and $\boldsymbol{w}$ is a

zero-mean normal random vector with variance $\boldsymbol{C}$, then the MVUE is

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x} \tag{2.6}$$

with estimator covariance equaling the CRB $\boldsymbol{I}^{-1}(\boldsymbol{\theta}) = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{-1}$.

Similarly, if $\boldsymbol{H}$ is not full rank, and $\boldsymbol{d}^T\boldsymbol{\theta}$ is an estimable function, then the MLE is $\boldsymbol{d}^T\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}(\boldsymbol{x})$ where $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}(\boldsymbol{x}) = \hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x})$ from (2.5). This MLE is also the MVUE [57, theorems 11.3F-G].

2.4 Maximum likelihood

Given observations $\boldsymbol{x}$ from a likelihood (or pdf) $p(\boldsymbol{x};\boldsymbol{\theta})$ depending on an unknown parameter $\boldsymbol{\theta}$, a popular method of estimating the parameter is the method of maximum likelihood. This approach chooses as an estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x})$ that, if true, would have the highest probability (the maximum likelihood) of resulting in the given observations $\boldsymbol{x}$, i.e., the optimization problem:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta})$$

where for convenience the log-likelihood is equivalently maximized since $\log(\cdot)$ is monotone. An analytic solution of the MLE can be found from the first-order conditions on the log-likelihood by considering solutions $\dot{\boldsymbol{\theta}}(\boldsymbol{x})$ of

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^{'}) = \boldsymbol{0}. \tag{2.7}$$

This is the *method of maximum likelihood*. Provided Θ is an open set, $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x})$ will satisfy (2.7).

2.4.1 Efficient estimation

If an efficient estimator exists, it is well-known that the method of maximum-likelihood finds the estimator [36, exercise 7.12], i.e., such an estimator must be a stationary point of the maximum-likelihood optimization problem. More formerly, we have the following theorem.

Theorem 2.8. If $\boldsymbol{t}(\boldsymbol{x})$ is an estimator of $\boldsymbol{\theta}$, which is also efficient with respect to the CRB, then the estimator is a stationary point of the following optimization problem:

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x}; \boldsymbol{\theta}).$$

2.4.2 Asymptotic Normality

Let the samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be iid as $\boldsymbol{x}$ from the pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$. Denote $\boldsymbol{y}_n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ to be the collection of these samples, so that the likelihood will be $p(\boldsymbol{y}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta})$, with the maximum likelihood of these samples denoted $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$.

Theorem 2.9. Assuming the regularity conditions stated earlier on the pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$, the MLE of the parameter $\boldsymbol{\theta}$ is asymptotically distributed according to

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})\right)$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is derived from the pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$, i.e., it is the Fisher information of a single observation or sample.

2.4.3 Scoring

There exists a number of approaches to finding the maximum likelihood, in some cases requiring iterative techniques. One such technique is Fisher's method of scoring. Given an observation $\boldsymbol{x}$ from a likelihood or pdf $p(\boldsymbol{x};\boldsymbol{\theta})$ depending on an unknown parameter $\boldsymbol{\theta}$ and given some initial estimate $\dot{\boldsymbol{\theta}}^{(1)}$ of $\boldsymbol{\theta}$, then iteratively using the update

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \boldsymbol{I}^{-1}(\dot{\boldsymbol{\theta}}^{(k)})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)}) \tag{2.8}$$

will find the MLE under certain conditions, e.g., provided the initial estimate is sufficiently close in a locally convex region.

2.5 Hypothesis testing

Given the inclusion of the CRB quantity in the asymptotic normality results in section 2.4.2, it is not surprising that there would also exist connections to some asymptotic hypothesis tests. Assume $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^r$ is a consistent and nonredundant differentiable function, which defines the null hypothesis

$$H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$$

in the likelihood (or model) $p(\boldsymbol{y}_n;\boldsymbol{\theta})$ versus the alternative hypothesis $H_1 : \boldsymbol{h}(\boldsymbol{\theta}) \neq \boldsymbol{0}$.

2.5.1 The Rao statistic

The Rao (or score) test statistic is given by

$$\rho(\boldsymbol{y}_n) = \frac{1}{n}\boldsymbol{s}^T(\boldsymbol{y}_n;\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{s}(\boldsymbol{y}_n;\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \tag{2.9}$$

where the Fisher score at $\boldsymbol{y}_n$ is $\boldsymbol{s}(\boldsymbol{y}_n, \hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) = \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))$ and $\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)$ is the MLE of $\boldsymbol{\theta}$ under the null hypothesis $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$. A variant of this statistic, the Lagrange multiplier test [3, 63]

$$\hat{\boldsymbol{\lambda}}_n^T \boldsymbol{H}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \hat{\boldsymbol{\lambda}}_n,$$

was developed by Silvey [63]. The equivalence between the Rao and Lagrange multiplier test comes from the first order condition to satisfy the constraint [45], i.e.,

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{y}_n; \hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) + \boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \hat{\boldsymbol{\lambda}}_n &= \boldsymbol{0}, \\ \boldsymbol{h}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) &= \boldsymbol{0} \end{aligned}$$

where $\hat{\boldsymbol{\lambda}}_n \in \mathbb{R}^r$ is a vector of Lagrange multiplier estimates. First order Taylor-series expansions of both equations about the true parameter $\boldsymbol{\theta}$ produces

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{y}_n; \boldsymbol{\theta}) - \boldsymbol{I}_n(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n) - \boldsymbol{\theta}) + \boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \hat{\boldsymbol{\lambda}}_n &= o(n^{-1/2}) \\ \boldsymbol{h}(\boldsymbol{\theta}) + \boldsymbol{H}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n) - \boldsymbol{\theta}) + o(n^{-1/2}) &= \boldsymbol{h}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \end{aligned}$$

where $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$ (n times the Fisher information based on a single sample $\boldsymbol{x}$). Hence under the null hypothesis, the latter implies $\boldsymbol{H}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n) - \boldsymbol{\theta}) = o(n^{-1/2})$, and premultiplying the former by $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$, then

$$\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{y}_n; \boldsymbol{\theta}) + \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \hat{\boldsymbol{\lambda}}_n = o(n^{-1/2}).$$

Since $\boldsymbol{s}(\boldsymbol{y}_n; \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n(\boldsymbol{\theta}))$, then applying Slutsky's theorem and the continuity of the Fisher information and the hypothesis function,

$$\boldsymbol{H}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)) \hat{\boldsymbol{\lambda}}_n \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta})\right).$$

Therefore, $\left(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\right)^{-1/2}\hat{\boldsymbol{\lambda}}_n$ is a r-dimensional standard normal variable in distribution, and

$$\rho(\boldsymbol{y}_n) \xrightarrow{d} \chi_r^2.$$

Hypothesis H_0 is rejected if $\rho(\boldsymbol{y}_n) > \chi_{r,\alpha}^2$, where $\chi_{r,\alpha}^2$ is the $(1-\alpha)$-quantile of the chi-square distribution with r degrees of freedom.

2.5.2 The Wald statistic

The Wald test statistic is given by

$$\omega(\boldsymbol{y}_n) = n\boldsymbol{h}^T(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\left(\boldsymbol{H}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\right)^{-1}\boldsymbol{h}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)) \tag{2.10}$$

where $\boldsymbol{H}(\boldsymbol{\theta}) = \left.\dfrac{\partial \boldsymbol{h}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}\right|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$ is an MLE of $\boldsymbol{\theta}$. (The nonredundancy of $\boldsymbol{h}$ implies that $\boldsymbol{H}(\cdot)$ is full row rank.) From section 2.1.1, the CRB of $\boldsymbol{h}(\boldsymbol{\theta})$ in the model $p(\boldsymbol{x};\boldsymbol{\theta})$ is $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta})$ and therefore, using theorem 2.9,

$$\sqrt{n}\left(\boldsymbol{h}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)) - \boldsymbol{h}(\boldsymbol{\theta})\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta})\right).$$

Hence, under H_0, using Slutsky's theorem, the convergence in probability of the MLE, and continuity of the FIM and Jacobian of the test function,

$$\omega(\boldsymbol{y}_n) \xrightarrow{d} \chi_r^2.$$

Therefore, the hypothesis H_0 is rejected if $\omega(\boldsymbol{y}_n) > \chi_{r,\alpha}^2$.

2.6 Discussion

In this section, the Cramér-Rao bound (CRB) was defined and in theorem 2.1 it was stated to be a bound on mean-square error performance of an unbiased

estimator. The theory of the CRB's connection to a variety of useful theorems and equations in mathematical statistics was demonstrated. The CRB is included in conditions for local identifiability (theorem 2.5) as well as conditions for strict identifiability (theorem 2.6). The CRB is the projection matrix of the BLUE under a Gaussian model in equation (2.6). There is a connection between the existence of efficiency with respect to the CRB and the method of ML (theorem 2.8). The CRB is also the asymptotic variance of the ML estimator (theorem 2.9) and appears in the update formula in (2.8) for the method of scoring. The CRB also appears in the formulas for the Rao test statistic in (2.9) and for the Wald test statistic (2.10).

These connections are not exhaustive, e.g., the CRB formula can also be useful in defining confidence regions or in useful as a cost function, but these are perhaps the most prevalent general topics in mathematical statistics theory and for that reason serve as a useful comparison for the *constrained* Cramér-Rao bound in chapter 3 and its connections in the theory.

Chapter 3

THE CONSTRAINED CRAMÉR-RAO BOUND

While the Cramér-Rao bound (CRB) is a useful measure of parametric estimation, it does not inherently measure the performance of estimators of parameters that satisfy side information in the form of a functional equality constraint

$$\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}. \tag{3.1}$$

The statistical literature is surprisingly somewhat limited in addressing performance measures under this general scenario. The traditional practice is to find some equivalent reparameterization of the particular model and then find the CRB on the parameters of interest using the reparameterized transformation. This approach, however, does not lend itself to theoretical meaning beyond the particular reparameterized model. Typically, works that do examine (3.1) in a general manner are focused on developing methods for decisions (hypothesis testing) instead of measuring estimation performance. Nevertheless, these results have connections to a CRB incorporating the side information in (3.1), or a constrained CRB. A number of papers, including Aitchison and Silvey [3] and Crowder [18], using the method of Lagrangian multipliers, examine the asymptotic normality of the constrained maximum likelihood estimator (CMLE) and as a consequence unintentionally develop a CRB under equality constraints. Under certain conditions, the asymptotic variance of the MLE equaling the CRB lends credence to the claim that the asymptotic vari-

ance of the CMLE should equal to the CRB under equality constraints, although the authors did not always appear cognizant of this fact. Others, including Silvey [63] to develop his Lagrange multiplier test, Osborne [55] with linear constraints to develop a scoring algorithm, and Waldorp, Huizenga, and Grasman [73] to develop a Wald-type test, also use the Lagrange multiplier approach in developing a constrained bound. Again, since these authors were primarily focused on asymptotic properties or hypothesis testing, the nature and perhaps utility of this mathematical quantity in their work is not explicitly stated as a CRB or bound on performance estimation of parameters under constraints. The creation of a constrained bound strictly for the use in performance analysis wasn't achieved until Gorman and Hero [23]. Gorman and Hero derived such a measure by taking the limit of the Hammersley-Chapman-Robbins bound with test points restricted to exist only in the constraint space. This constrained Cramér-Rao bound (CCRB)

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^{T}(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^{T}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \tag{3.2}$$

utilizes the Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$ of the functional constraint $\boldsymbol{f}(\boldsymbol{\theta})$ and the inverse of the Fisher Information matrix (FIM) $\boldsymbol{I}(\boldsymbol{\theta})$ (based on the unconstrained model), which must be nonsingular. As with the CRB, there exist a number of alternative derivations. The works of Crowder, Waldorp, et al, Gorman and Hero, and Aitchison [2] include the formula in (3.2) in some manner for which the CRB might be used for the unconstrained scenario in their works, a fact that implicitly proves the validity of the CCRB. With the explicit proof by Gorman and Hero as a guideline, Marzetta [47] provides an elementary proof of this CCRB, which avoids the use of the application of

the Cauchy-Schwarz inequality, and avoids the use of pseudo-inverses, by examining the inequality created from the positive-semidefiniteness property for the variance of a properly defined random variable. A similar construction was used by Stoica and Ng [68] to formulate a more general CCRB

$$\boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta}) \tag{3.3}$$

that incorporates the constraint information without the assumption of a nonsingular FIM. This CCRB utilizes the unconstrained FIM and an orthonormal complement matrix $\boldsymbol{U}(\boldsymbol{\theta})$ whose vectors span the null space of the constraint Jacobian matrix. Furthermore, when the FIM is nonsingular, the Stoica-Ng CCRB in (3.3) is equivalent to the Gorman-Hero version of (3.2). Hence, while much of the previous work used the formula in (3.2), the more general formula (3.3) is also applicable. Osborne, independently from much of these other works, developed a method of scoring with constraints that utilizes the Stoica-Ng CCRB formula in (3.3) as the projection matrix in place of the CRB. There are, of course, numerous instances of matrix structures of the same form as (3.3), for example as part of the projection matrix of the generalized least squares estimate of the mean of a linear model.

In this section, we develop a very simple derivation of the CCRB in (3.3). Rather than assuming the parameters satisfy functional constraints, we approach the problem theoretically from an alternative, yet equivalent, perspective and assume the parameters locally fit a reduced parametric model. This approach permits the extension of the existing classical theory underlying the CRB to the case of a model under parametric constraints. While it is true that several of these exten-

sions already exist in the literature, there does not exist a cohesive treatment of these results. However, this chapter should not be viewed as simply a collection of historical results, but a unified and comprehensive development of the theory of the constrained Cramér-Rao bound.

3.1 The Constrained CRB

Suppose we observe $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^n$ from a probability density function $p(\boldsymbol{x}; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ is a vector of unknown deterministic parameters and, in addition, suppose these parameters are required to satisfy k *consistent* and *nonredundant* continuously differentiable parametric equality constraints, i.e., $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$ for some consistent and nonredundant $\boldsymbol{f} : \Theta \to \mathbb{R}^k$. We shall denote

$$\Theta_f = \left\{ \boldsymbol{\theta}' \in \Theta : \boldsymbol{f}(\boldsymbol{\theta}') = \mathbf{0}, \boldsymbol{f} \text{ consistent, nonredundant} \right\} \tag{3.4}$$

to be the *feasible* set satisfying the constraints. Hence, the constraint can also be stated $\boldsymbol{\theta} \in \Theta_f$. The constraints being consistent means that the set Θ_f is nonempty. The constraints being nonredundant means that the Jacobian $\boldsymbol{F}(\boldsymbol{\theta}') = \frac{\boldsymbol{f}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}$ has rank k whenever $\boldsymbol{f}(\boldsymbol{\theta}') = \mathbf{0}$.

As before, the Fisher information matrix (FIM) of this model (ignoring the constraint) is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left\{ \boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta}) \boldsymbol{s}^T(\boldsymbol{x}; \boldsymbol{\theta}) \right\}$$

where $\boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta})$ is the Fisher score defined by

$$\boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta}) = \left. \frac{\partial \log p(\boldsymbol{x}; \boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}' = \boldsymbol{\theta}}$$

and the expectation is evaluated at $\boldsymbol{\theta}$, i.e., $E_{\boldsymbol{\theta}}(\cdot) = \int_{\mathbb{X}} (\cdot) p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}$.

Incorporating this additional *side information* into the information from the observations $\boldsymbol{x}$ directly would require an alteration of the pdf's dependence on the unknown parameter. Such an approach can often be analytically impractical or numerically complex. Hence, it is desirable to have a formulaic or prescriptive approach to include the side information, or constraints, indirectly. To meet this need, Stoica and Ng developed a method to incorporate parametric equality constraints into the CRB [68, theorem 1].

Theorem 3.1 (Stoica & Ng)**.** The constrained Cramér-Rao bound on $\boldsymbol{\theta} \in \Theta_f$ is given by

$$\mathrm{CCRB}(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta}) \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{U}(\boldsymbol{\theta}) \right)^{-1} \boldsymbol{U}^T(\boldsymbol{\theta}) \tag{3.5}$$

where $\boldsymbol{U}(\boldsymbol{\theta})$ is a matrix whose column vectors form an orthonormal basis for the null space of the Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{0} \ , \ \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{I}_{(m-k)\times(m-k)}. \tag{3.6}$$

Thus, if $\boldsymbol{t}(\boldsymbol{x})$ is an unbiased estimator of $\boldsymbol{\theta}$, which satisfies the constraint (3.4), then

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \mathrm{CCRB}(\boldsymbol{\theta})$$

with equality if and only if $\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} = \mathrm{CCRB}(\boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ in the mean-square sense.[1]

[1] The original theorem requires the estimator to satisfy the constraint. In general, the parameter and its unbiased estimator will not simultaneously satisfy the constraint since the implication, mainly that $\boldsymbol{f}(E_{\boldsymbol{\theta}}\boldsymbol{t}(\boldsymbol{x})) = E_{\boldsymbol{\theta}}\boldsymbol{f}(\boldsymbol{t}(\boldsymbol{x}))$, is true only under particular conditions. However, the CCRB is the same if either assumption is made exclusively. In this treatise, I assume that the actual parameter $\boldsymbol{\theta}$ satisfies the constraint and the unbiased estimator $\boldsymbol{t}(\boldsymbol{x})$ does not. (In section 3.4, the constrained maximum likelihood estimator (CMLE) is assumed to satisfy the constraint, but unbiasedness is not assumed.)

The Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$ having full row rank is not necessary since the Jacobian does not explicitly appear in the CCRB formula in (3.5). Indeed, the requirement that the column vectors of $\boldsymbol{U}(\boldsymbol{\theta})$ form an orthonormal basis is also unnecessary, only that they be linearly independent and that they span the basis of the null space of the row vectors of $\boldsymbol{F}(\boldsymbol{\theta})$, i.e., only that the column space of $\boldsymbol{U}(\boldsymbol{\theta})$ is an *orthogonal complement* of the row space of $\boldsymbol{F}(\boldsymbol{\theta})$, is required since it is clear from the structure of (3.5) that the CCRB is invariant to automorphisms in $\mathbb{R}^{m-\operatorname{rank}(\boldsymbol{F}(\boldsymbol{\theta}))}$ on $\boldsymbol{U}(\boldsymbol{\theta})$. Regardless, for convenience, and except where otherwise noted, we will assume that the constraints are nonredundant and the columns of $\boldsymbol{U}(\boldsymbol{\theta})$ are orthonormal to ensure that $\operatorname{rank}(\boldsymbol{U}(\boldsymbol{\theta})) = m - k$ and $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{I}_{m-k}$, respectively. The existence of the bound only requires that $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ rather than the FIM be nonsingular.[2] The original proof of this theorem given by Stoica and Ng considers the variance inequality generated by the random variable $\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} - \boldsymbol{W}\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ and maximizes $\boldsymbol{W}$ to attain the tightest bound for $\operatorname{Var}(\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta})$.

Example 3.2 (Unit Modulus Constraint)**.** Suppose ϑ in example 2.2 is constrained to be unit modulus, i.e., $f(\vartheta) = |\vartheta|^2 - 1$. Then its gradient in terms of the parameter vector $\boldsymbol{\theta} = \begin{bmatrix}\operatorname{Re}(\vartheta)\\ \operatorname{Im}(\vartheta)\end{bmatrix}$ is $\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix}2\operatorname{Re}(\vartheta),\ 2\operatorname{Im}(\vartheta)\end{bmatrix}$, which has an orthonormal complement $\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix}-\operatorname{Im}(\vartheta)\\ \operatorname{Re}(\vartheta)\end{bmatrix}$. The CCRB for this constraint is then

$$\operatorname{CCRB}(\boldsymbol{\theta}) = \frac{\sigma^2}{2}\begin{bmatrix}\operatorname{Im}(\vartheta)^2 & -\operatorname{Im}(\vartheta)\operatorname{Re}(\vartheta)\\ -\operatorname{Re}(\vartheta)\operatorname{Im}(\vartheta) & \operatorname{Re}(\vartheta)^2\end{bmatrix}.$$

[2] A corresponding regularity condition to that mentioned in section 2.1 will be discussed in section 3.1.4.

3.1.1 A proof of the CCRB

While the proof given by Stoica and Ng is sufficient to establish the validity of the CCRB, these proofs ignore the existing classical theory encompassing the CRB and FIM, which is already sufficient to prove the CCRB. However, prior to developing the foundation for the CCRB from the existing theory, we require a foray into multivariable calculus and, specifically, the use of the implicit function theorem. The reward for this approach will be a seamless presentation of statistical inference involving the constrained Cramér-Rao bound.

From the perspective of multivariable calculus, the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$ effectively restricts $\boldsymbol{\theta}$ to a manifold Θ_f of the original vector space Θ, with the manifold having dimension $m - k$ since k degrees of freedom are lost when $\text{rank}(\boldsymbol{F}(\boldsymbol{\theta})) = k$ for all $\boldsymbol{\theta} \in \Theta_f$. More formally, the following theorem [65, theorems 5-1 and 5-2] applies.

Theorem 3.3 (Implicit Function Theorem)**.** Let $\mathbb{U} \subset \mathbb{R}^m$ be an open set and assume $\boldsymbol{f} : \mathbb{U} \to \mathbb{R}^k$ is a differentiable function such that $\boldsymbol{F}(\boldsymbol{\theta})$ has rank k whenever $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$. Then $\Theta_f \cap \mathbb{U}$ is an $(m-k)$-dimensional manifold in $\mathbb{R}^m$, and for every $\boldsymbol{\theta} \in \Theta_f \cap \mathbb{U}$ there is an open set $\mathbb{V} \ni \boldsymbol{\theta}$, an open set $\mathbb{W} \subset \mathbb{R}^{m-k}$, and a 1-1 differentiable function $\boldsymbol{g}_{\boldsymbol{\theta}} : \mathbb{W} \to \mathbb{R}^m$ such that

(a) $\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbb{W}) = \Theta_f \cap \mathbb{U} \cap \mathbb{V}$, and

(b) the Jacobian of $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}')$ has rank $m - k$ for each $\boldsymbol{\xi}' \in \mathbb{W}$.

Therefore, there exists a function $\boldsymbol{g}_{\boldsymbol{\theta}} : \mathbb{R}^{m-k} \to \mathbb{R}^m$, and sets $\mathbb{O} \ni \boldsymbol{\theta}$ and $\mathbb{P}$ open in Θ_f and $\mathbb{R}^{m-k}$, respectively, such that $\boldsymbol{g}_{\boldsymbol{\theta}} : \mathbb{P} \to \mathbb{O}$ is a *diffeomorphism* on $\mathbb{P}$,

i.e., a continuously differentiable bijection with a continuously differentiable inverse. A geometric example is shown in Figure 3.1. Note this diffeomorphism depends on the parameter $\boldsymbol{\theta}$ as the reparameterization is only guaranteed to exist in a local neighborhood of $\boldsymbol{\theta}$; however, for convenience, we will omit this notation in this subsection so that $\boldsymbol{g} = \boldsymbol{g}_{\boldsymbol{\theta}}$. Thus, we may proceed under the assumption that every $\boldsymbol{\theta}' \in \mathbb{O} \subset \Theta_f$ is the image of a unique reduced parameter vector $\boldsymbol{\xi}' \in \mathbb{P} \subset \mathbb{R}^{m-k}$ under $\boldsymbol{g}$, or simply

$$\boldsymbol{\theta}' = \boldsymbol{g}(\boldsymbol{\xi}'). \tag{3.7}$$

Necessarily, there exists some unique $\boldsymbol{\xi} \in \mathbb{P}$ for which $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$. We will denote the Jacobian of $\boldsymbol{g}$ to be $\boldsymbol{G}(\boldsymbol{\xi}') = \frac{\partial \boldsymbol{g}(\boldsymbol{\xi}')}{\partial \boldsymbol{\xi}'^T}$, which also implicitly depends on $\boldsymbol{\theta}$.

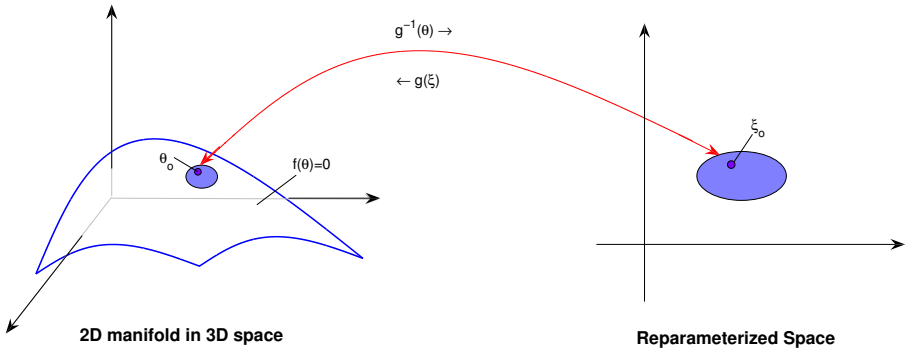


Figure 3.1: Reparameterization of $\boldsymbol{f}(\boldsymbol{\theta}) = 0$ to $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$.

Example 3.4 (Unit Modulus Constraint)**.** As an example of this principle, consider a complex parameter ϑ with a modulus constraint (as in example 3.2). The parameter vector in this case may be $\boldsymbol{\theta} = [\theta_1, \theta_2]^T = [\mathrm{Re}(\vartheta), \mathrm{Im}(\vartheta)]^T \in \mathbb{R}^2$ with the constraint being $f(\boldsymbol{\theta}) = \theta_1^2 + \theta_2^2 - 1 = 0$. By the implicit function theorem,

constraining ϑ to be unit modulus is tantamount to assuming the existence of a $\xi \in \mathbb{R}$ such that $\vartheta = e^{-j\xi}$. Hence, $\boldsymbol{\theta}$ is a function of ξ, i.e.,

$$\boldsymbol{\theta} = \begin{bmatrix} \mathrm{Re}(\vartheta) \\ \mathrm{Im}(\vartheta) \end{bmatrix} = \begin{bmatrix} \cos(\xi) \\ -\sin(\xi) \end{bmatrix} = \boldsymbol{g}(\xi). \tag{3.8}$$

Also, $\boldsymbol{g}(\mathbb{R}) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : f(\boldsymbol{\theta}) = 0\}$ and $\boldsymbol{G}(\xi) = \begin{bmatrix} -\sin(\xi) & -\cos(\xi) \end{bmatrix}^T$ has rank 1 for every ξ. For the model in example 3.2, then

$$\mathrm{CCRB}(\boldsymbol{\theta}) = \frac{\sigma^2}{2} \begin{bmatrix} \sin^2(\xi) & -\sin(\xi)\cos(\xi) \\ -\cos(\xi)\sin(\xi) & \cos^2(\xi) \end{bmatrix},$$

which is exactly as before.

It must be noted that $\boldsymbol{g}$ will not be unique. For the previous example, $\vartheta = e^{j\xi}$ is another possible reparameterization. Nor is any $\boldsymbol{g}$ satisfying the theorem necessarily a 1-1 correspondence between $\mathbb{R}^{m-k}$ and Θ_f; again, for the previous example, $\boldsymbol{g}$ is periodic. Thus, the bijection is only guaranteed locally. Finding a particular $\boldsymbol{g}$ for a given $\boldsymbol{f}$ and $\boldsymbol{\theta}$ may not be obvious. Methods for approximating an implicit function will be discussed in section 3.1.5. Regardless, as shall be shown in the context of the CCRB, knowledge of any particular $\boldsymbol{g}$ is unnecessary; only its existence, given by the implicit function theorem, is necessary. Why? Using the implicit function theorem to assume a locally equivalent reparameterization for the constraint limits the information from the observations to the density function's local dependence on the unknown parameter. But as the CRB (and hence CCRB) is a local bound that only characterizes the local noise ambiguities in the model, i.e., the average local curvature of the density at the parameter value of interest, this limitation is invariant to the local curvature restricted to a space determined by the constraints.

Theorem 3.5 ([50])**.** The CRB on $\boldsymbol{\theta} \in \mathbb{O}$ under the assumption of (3.7) is given by

$$\boldsymbol{G}(\boldsymbol{\xi}) \left(\boldsymbol{G}^T(\boldsymbol{\xi}) \boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\xi})) \boldsymbol{G}(\boldsymbol{\xi}) \right)^{-1} \boldsymbol{G}^T(\boldsymbol{\xi}) \tag{3.9}$$

and if $\boldsymbol{g}$ in (3.7) is an implicit function of $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$ then this bound is equivalent to the constrained Cramér-Rao bound in (3.5).

Proof. This CRB can be developed from a transformation of parameters on the FIM and the CRB. From the CRB transformation of parameters on the CRB (see section 2.1.1), if $\boldsymbol{t}(\boldsymbol{x})$ is an unbiased estimator[3] of $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$ then

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \mathrm{CRB}(\boldsymbol{g}(\boldsymbol{\xi})) = \boldsymbol{G}(\boldsymbol{\xi}) \tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi}) \boldsymbol{G}^T(\boldsymbol{\xi}) \tag{3.10}$$

with equality if and only if $\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\xi}) = \boldsymbol{G}(\boldsymbol{\xi}) \tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi}) \tilde{\boldsymbol{s}}(\boldsymbol{x}; \boldsymbol{\xi})$ in the mean-square sense [36, Appendix 3B], where $\tilde{\boldsymbol{I}}(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}} \tilde{\boldsymbol{s}}(\boldsymbol{x}; \boldsymbol{\xi}) \tilde{\boldsymbol{s}}^T(\boldsymbol{x}; \boldsymbol{\xi})$ is the FIM on $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{s}}(\boldsymbol{x}; \boldsymbol{\xi}) = \left. \frac{\partial \log q(\boldsymbol{x}; \boldsymbol{\xi}')}{\partial \boldsymbol{\xi}'} \right|_{\boldsymbol{\xi}' = \boldsymbol{\xi}}$ is the Fisher score of the pdf with respect to $\boldsymbol{\xi}$, this pdf being $q(\boldsymbol{x}; \boldsymbol{\xi}) = p(\boldsymbol{x}; \boldsymbol{g}(\boldsymbol{\xi}))$. By application of the derivative chain rule the Fisher score of $\boldsymbol{x}$ with respect to $\boldsymbol{\xi}$ is $\tilde{\boldsymbol{s}}(\boldsymbol{x}; \boldsymbol{\xi}) = \boldsymbol{G}^T(\boldsymbol{\xi}) \boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta})$. Hence, from the transformation of parameters on the FIM [12, p. 157], the FIM on $\boldsymbol{\xi}$ is[4]

$$\begin{aligned} \tilde{\boldsymbol{I}}(\boldsymbol{\xi}) &= E\left\{ \tilde{\boldsymbol{s}}(\boldsymbol{x}; \boldsymbol{\xi}) \tilde{\boldsymbol{s}}^T(\boldsymbol{x}; \boldsymbol{\xi}) \right\} \\ &= E\left\{ \boldsymbol{G}^T(\boldsymbol{\xi}) \boldsymbol{s}(\boldsymbol{x}; \boldsymbol{\theta}) \boldsymbol{s}^T(\boldsymbol{x}; \boldsymbol{\theta}) \boldsymbol{G}(\boldsymbol{\xi}) \right\} \\ &= \boldsymbol{G}^T(\boldsymbol{\xi}) \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{G}(\boldsymbol{\xi}) \\ &= \boldsymbol{G}^T(\boldsymbol{\xi}) \boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\xi})) \boldsymbol{G}(\boldsymbol{\xi}). \end{aligned} \tag{3.11}$$

[3] Again, there is no actual use of the assumption here that $\boldsymbol{t}(\boldsymbol{x}) \in \Theta_f$ although if $\boldsymbol{t}(\boldsymbol{x}) \in g(\mathcal{P})$ then $\boldsymbol{t}(\boldsymbol{x})$ does indeed satisfy the constraint. The theorem result for unbiased estimators holds regardless and this will depend on the regularity condition detailed in section 3.1.4.

[4] Implicitly, it is assumed that the regularity conditions of section 2.1 hold with respect to $\boldsymbol{\xi}$. For how this applies to $\boldsymbol{\theta}$, see section 3.1.4.

Substituting (3.11) into (3.10) establishes the CRB under the assumption $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$. To establish the equivalence with the CCRB, note since $\boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{\xi}) = \mathbf{0}$, then by taking the Jacobian with respect to $\boldsymbol{\xi}$ we have

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\xi}^T}\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\xi}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{\xi})) = \boldsymbol{F}(\boldsymbol{g}(\boldsymbol{\xi}))\boldsymbol{G}(\boldsymbol{\xi}) = \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta})).$$

Hence, the columns of $\boldsymbol{G}(\boldsymbol{\xi})$ reside in the null space of the row vectors of $\boldsymbol{F}(\boldsymbol{\theta})$. And since $\boldsymbol{g}$ has an inverse locally at $\boldsymbol{\xi} = \boldsymbol{g}^{-1}(\boldsymbol{\theta})$, then $\boldsymbol{G}(\boldsymbol{\xi})$ has full column rank $m - k$ locally about $\boldsymbol{\xi}$. (This is true on the whole set $\mathbb{P} \subset \mathbb{R}^{m-k}$.) Therefore, $\text{span}(\boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta}))) = \text{span}(\boldsymbol{U}(\boldsymbol{\theta}))$ and there exists an $m - k \times m - k$ full rank transformation matrix $\boldsymbol{S}(\boldsymbol{\theta})$ such that $\boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta}))\boldsymbol{S}(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta})$. (This is true on the whole set $\mathbb{O} \subset \Theta_f$.) This matrix $\boldsymbol{S}(\boldsymbol{\theta})$ is merely an orthonormalizing change of basis on the columns of $\boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta}))$. Therefore,

$$\begin{aligned}
&\boldsymbol{G}(\boldsymbol{\xi})\left(\boldsymbol{G}^T(\boldsymbol{\xi})\boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\xi}))\boldsymbol{G}(\boldsymbol{\xi})\right)^{-1}\boldsymbol{G}^T(\boldsymbol{\xi}) \\
&= \boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta}))\left(\boldsymbol{S}^{-T}(\boldsymbol{\theta})\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{S}^{-1}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{G}^T(\boldsymbol{g}^{-1}(\boldsymbol{\theta})) \\
&= \boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{S}^T(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{g}^{-1}(\boldsymbol{\theta})) \\
&= \boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta}).
\end{aligned}$$

Also, $\text{Var}(\boldsymbol{t}(\boldsymbol{\xi})) \geq \text{CCRB}(\boldsymbol{\theta})$ with equality if and only if

$$\begin{aligned}
\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\xi}) &= \boldsymbol{G}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi})\tilde{\boldsymbol{s}}(\boldsymbol{x};\boldsymbol{\xi}) \\
\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} &= \boldsymbol{G}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi})\boldsymbol{G}^T(\boldsymbol{\xi})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}) \\
&= \text{CCRB}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}).
\end{aligned}$$

□

This proof of the CCRB uses the implicit function theorem, the CRB transformation formula, the FIM transformation formula, as well as well-known properties of rank and the derivative chain rule. An advantage of establishing the CCRB from these classical results will become clear as we establish the connection throughout statistical information theory. An example of this is immediately evident in an alternative proof of a proposition of Stoica and Ng [68, proposition 1].

Corollary 3.6 (Stoica & Ng)**.** Given the regularity conditions on $\boldsymbol{\xi}$, a necessary and sufficient condition for the existence of a finite CCRB of $\boldsymbol{\theta}$ is

$$\left|\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right| \neq 0,$$

i.e., $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is nonsingular.

Proof. From the prior theorem, it is clear that $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is nonsingular if and only if $\boldsymbol{G}^T(\boldsymbol{\xi})\boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\xi}))\boldsymbol{G}(\boldsymbol{\xi})$ is nonsingular if and only if $\tilde{\boldsymbol{I}}(\boldsymbol{\xi})$ is nonsingular. Since a necessary and sufficient condition for the existence of a finite CRB of $\boldsymbol{g}(\boldsymbol{\xi})$ is that $\tilde{\boldsymbol{I}}(\boldsymbol{\xi})$ is nonsingular, the corollary is proven. □

Thus, with the usual regularity conditions (see section 2.1) being maintained for $\boldsymbol{\xi}$, the conditions for the existence of the CCRB with respect to $\boldsymbol{\theta}$ are equivalent to the conditions for the existence of the CRB in the reduced parameter space of $\boldsymbol{\xi}$.

Moreover, the matrix $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is nonsingular if and only if $\boldsymbol{U}(\boldsymbol{\theta})$ has full column rank $m-k$ and no linear combination of the columns of $\boldsymbol{U}(\boldsymbol{\theta})$ reside in the null space of $\boldsymbol{I}(\boldsymbol{\theta})$. The first condition is satisfied always by definition. The second condition implies, for example, that if $\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}^T(\boldsymbol{\theta})$ is the Cholesky decomposition [20,

p.194] of the FIM, where $\boldsymbol{L}(\boldsymbol{\theta}) \in \mathbb{R}^{m \times \operatorname{rank}(\boldsymbol{I}(\boldsymbol{\theta}))}$ is a full column rank lower triangular matrix with strictly positive values on the diagonal, then $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta}) + \boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}(\boldsymbol{\theta})$ where $\boldsymbol{K}(\boldsymbol{\theta})$ is an orthogonal complement of $\boldsymbol{L}(\boldsymbol{\theta})$ and for some full column rank $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathbb{R}^{\operatorname{rank}(\boldsymbol{I}(\boldsymbol{\theta})) \times m-k}$.

Further properties of the CCRB will be discussed in section 3.1.4.

3.1.2 Alternative formulas

The formula Stoica and Ng used to express the constrained Cramér-Rao bound is a generalization of an expression developed earlier first by Gorman and Hero [23, theorem 1] and later by Marzetta [47, theorem 2]. This is the CCRB formula in (3.12). Although Gorman and Hero's formula requires a nonsingular Fisher information, the version developed by Stoica and Ng appears to be inspired by a result [23, (19) in lemma 2] in Gorman and Hero's paper that unnecessarily assumes a positive definite FIM. However, this result was, in essence, not unknown in the literature. Gorman and Hero were aware of the prior work of Aitchison and Silvey [3, theorem 2 and $\boldsymbol{P}$ on p.823], which is concerned with the asymptotic variance of the maximum likelihood estimator subject to restraints. But they were perhaps unaware (by lack of citation) of a later paper on hypothesis tests associated with the maximum likelihood, in which Aitchison and Silvey suggest a solution to the problem of singular information matrices [4, section 6]. This is the CCRB formula in (3.14). This version of the CCRB was also proven by Crowder [18, theorem 3]. Concurrent to the Gorman and Hero effort, Hendriks [27] and later Oller and Corcuera [53] developed

an extension of the Cramér-Rao bound intrinsic to the manifold using Riemannian geometry. More recently, Xavier and Barroso [74, 75] specified the lower bound on the geodesic of estimators to the true parameter. The original and latest versions of their bound are expressed in (3.16) and (3.17), respectively.

While these CCRB expressions are not the focus of the current treatise, they are still important for possible insights into the constrained Cramér-Rao bound. In this section, aspects of these insights will be briefly discussed as well as conditions for equality with the CCRB expression in (3.5).

3.1.2.1 Gorman-Hero-Aitchison-Silvey CCRB

Aitchison and Silvey used the method of Lagrange multipliers to show that the weighted asymptotic variance of the constrained maximum likelihood estimator, which should be the CCRB (implicitly), tends to

$$\mathrm{CCRB}_2(\boldsymbol{\theta}) = \boldsymbol{I}^{-1}(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \quad (3.12)$$

when the Fisher information is nonsingular. Alternatively, Gorman and Hero developed this same CCRB by restricting *test points* in the Chapman-Robbins bound (a Barankin-type bound) to be in the constraint space Θ_f and then finding the derivatives as the limit of the finite difference expressions in the Chapman-Robbins bound.[5] A simpler proof was provided by Marzetta by considering the positive semidefiniteness of a properly chosen random variable.

A particular advantage to this presentation of the CCRB is the explicit quan-

[5] A definition of the Chapman-Robbins bound as well as a variant of the Gorman and Hero proof, which allows for a singular FIM, can be found in Appendix A.1.

tification of the gain in performance potential. Imposing a constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$ on a set of parameters improves (lowers) the unconstrained bound from $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ by exactly $\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})$. Since the CRB of $\boldsymbol{f}(\boldsymbol{\theta})$ is $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})$ then its inverse is the Fisher information of $\boldsymbol{f}(\boldsymbol{\theta})$ or $\mathbf{0}$. And $\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})$ is the Fisher information of $\boldsymbol{\theta}$ generated from the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$. A disadvantage is the requirement of a nonsingular FIM. There exist numerous scenarios that require constraints for the original model to be identifiable (see Chapter 4). Additionally, this CCRB formula requires nonredundant constraints, i.e., the Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$ must be full row rank.

Similarities include the computational complexity of both formulas, which is $O(m^3)$. And when the Fisher information is nonsingular (and the constraints nonredundant), both formulas are equivalent.

Theorem 3.7. When the Fisher information is nonsingular and the constraints nonredundant, then an equivalent formula for the CCRB in (3.5) is $\text{CCRB}_2(\boldsymbol{\theta})$.

Proof. This is a different proof than the one provided in [68, corollary 1].[6] The existence of the Gorman-Hero-Marzetta formula assumes that the FIM $\boldsymbol{I}(\boldsymbol{\theta})$ and the CRB of the constraint $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})$ are regular (non-singular) [23, 47]. Correspondingly, the existence of the Stoica and Ng CCRB formula assumes that

[6] In addition to this alternate proof, they reference an algebraic identity from [37] that is useful in establishing the result.

Lemma 3.8 (Khatri)**.** Suppose $\boldsymbol{A}$ is $p \times q$ and $\boldsymbol{B}$ is $p \times p - q$ have ranks q and $p - q$ respectively such that $\boldsymbol{B}^T\boldsymbol{A} = \mathbf{0}$. Then for any symmetric positive definite matrix $\boldsymbol{S}$,

$$\boldsymbol{S}^{-1} - \boldsymbol{S}^{-1}\boldsymbol{A}\left(\boldsymbol{A}^T\boldsymbol{S}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{S}^{-1} = \boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{S}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T.$$

Substituting $\boldsymbol{I}(\boldsymbol{\theta})$ for $\boldsymbol{S}$, $\boldsymbol{F}(\boldsymbol{\theta})$ for $\boldsymbol{A}^T$, and $\boldsymbol{U}(\boldsymbol{\theta})$ for $\boldsymbol{B}$ shows the equivalence between the two CCRBs for when the FIM is nonsingular.

$\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is non-singular. Now, right-multiplying both formulas by $\boldsymbol{F}^T(\boldsymbol{\theta})$ returns the results

$$\begin{aligned}\mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) &= \boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) = \mathbf{0},\\ \mathrm{CCRB}_2(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) &= \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) = \mathbf{0}\end{aligned}$$

since $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ in the first equation and the elimination of the inverse in $\mathrm{CRB}(\boldsymbol{f}(\boldsymbol{\theta}))$ in the second. Alternatively, right-multiplying both formulas by the matrix $\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ returns the results

$$\begin{aligned}\mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) &= \boldsymbol{U}(\boldsymbol{\theta}),\\ \mathrm{CCRB}_2(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) &= \boldsymbol{U}(\boldsymbol{\theta})\end{aligned}$$

again by eliminating the inverse in the first equation and since $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ in the second. Hence we have the equality

$$\mathrm{CCRB}(\boldsymbol{\theta})\begin{bmatrix}\boldsymbol{F}^T(\boldsymbol{\theta}) & \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\end{bmatrix} = \mathrm{CCRB}_2(\boldsymbol{\theta})\begin{bmatrix}\boldsymbol{F}^T(\boldsymbol{\theta}) & \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\end{bmatrix}$$

and if it can be shown that the matrix $\begin{bmatrix}\boldsymbol{F}^T(\boldsymbol{\theta}), & \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\end{bmatrix}$ is regular, then the two CCRB formulas are shown to be equivalent. Suppose there exists vectors $\boldsymbol{\alpha} \in \mathbb{R}^k$ and $\boldsymbol{\beta} \in \mathbb{R}^{m-k}$ such that

$$\boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{\alpha} + \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{\beta} = \mathbf{0}. \tag{3.13}$$

Pre-multiplying (3.13) by $\boldsymbol{U}^T(\boldsymbol{\theta})$ implies that

$$\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{\beta} = \mathbf{0}.$$

Since $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is regular then $\boldsymbol{\beta} = \mathbf{0}$. Likewise, premultiplying (3.13) by $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ implies that

$$\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{\alpha} = \mathbf{0}.$$

Since $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})$ is regular then $\boldsymbol{\alpha} = \mathbf{0}$. Hence $\boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{\alpha} + \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{\beta} = \mathbf{0}$ implies $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{0}$, which proves that $\begin{bmatrix}\boldsymbol{F}^T(\boldsymbol{\theta}) & \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\end{bmatrix}$ is full rank. □

3.1.2.2 Aitchison-Silvey-Crowder CCRB

The solution for resolving invertibility of singular FIMs in the variance and test results of Aitchison and Silvey [4] was to load the Fisher information with a matrix of the form $\boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$. This was made more rigorous by Crowder [18] by replacing the Fisher information $\boldsymbol{I}(\boldsymbol{\theta})$ with a *loaded FIM* $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{K}\boldsymbol{F}(\boldsymbol{\theta})$, where $\boldsymbol{K}$ is any positive semidefinite matrix such that $\boldsymbol{D}(\boldsymbol{\theta})$ is regular. Hence, a generalization of (3.12) is

$$\mathrm{CCRB}_3(\boldsymbol{\theta}) = \boldsymbol{D}^{-1}(\boldsymbol{\theta}) - \boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta}). \tag{3.14}$$

This extension permits a singular FIM. This formulation is also independent of the choice of $\boldsymbol{K}$.

Theorem 3.9. An equivalent formula for the CCRB in (3.5) is $\mathrm{CCRB}_3(\boldsymbol{\theta})$.

Proof. Replace $\boldsymbol{I}(\boldsymbol{\theta})$ with $\boldsymbol{D}(\boldsymbol{\theta})$ in the proof of theorem 3.7.[7] □

[7] In appendix A.3, this Crowder formula for the asymptotic variance of the constrained maximum likelihood estimate is shown to be equivalent to the CCRB using lemma 3.8 (also see section 3.4.2).

As the computational complexity of the two formulas are $O(m^3)$, there is no direct advantage of one CCRB version over the other. Certainly, in the nonsingular FIM case, $\text{CCRB}_3(\boldsymbol{\theta})$ is the same as the $\text{CCRB}_2(\boldsymbol{\theta})$ (with $\boldsymbol{K} = \mathbf{0}$). Unfortunately, this CCRB formula does not appear to have simple connections to other areas of statistical inference in an inherent manner. Nevertheless, this possible inefficacy in theoretical applications does not effect its practical use.

3.1.2.3 Xavier's & Barroso's Intrinsic Variance Lower Bound

The prior CCRB metrics were in Euclidean $\mathbb{R}^m$ space, i.e., the lower bound is on the measurement of the distance (in some direction or dimension) of the estimator to the true value of the parameter measured by "cutting through" the manifold. In some scenarios, it may be of interest to know what the bound is on the measurement of the distance "over the surface" of the manifold. Since dimensional directions can be somewhat ambiguous depending on the manifold, of particular interest is the geodesic, or shortest distance.

For this scenario, Xavier and Barroso [74, 75] formulated an inequality, the intrinsic variance lower bound (IVLB), for the variance of the geodesic to an unbiased estimator intrinsic to the manifold. Their results are derived from those of Hendriks [27] and Oller and Corcuera [53]. Ignoring elements of Riemannian geometric theory that are beyond the scope of this presentation, their IVLB result essentially relies on the inequality

$$\sqrt{C}\sqrt{\text{var}(\vartheta)}\cot(\sqrt{C}\sqrt{\text{var}(\vartheta)}) \leq \frac{\sqrt{\text{var}(\vartheta)}}{\sqrt{\lambda_{\boldsymbol{\theta}}}}, \tag{3.15}$$

where C is an upper bound on the sectional curvature of the manifold, $\lambda_{\boldsymbol{\theta}}$ is a bound on variance of the Euclidean estimator error, and $\text{var}(\vartheta)$ is the variance of the geodesic. Precisely, $C = \max_{\boldsymbol{\theta}\in\Theta_f} K_{\boldsymbol{\theta}}$ where

$$K_\theta = \max_{\substack{\boldsymbol{v}_1,\boldsymbol{v}_2 \text{ orthonormal} \\ \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{v}_i=\boldsymbol{0}}} < \Pi(\boldsymbol{v}_1,\boldsymbol{v}_1), \Pi(\boldsymbol{v}_2,\boldsymbol{v}_2) > - < \Pi(\boldsymbol{v}_1,\boldsymbol{v}_2), \Pi(\boldsymbol{v}_1,\boldsymbol{v}_2) >$$

and $\Pi(\cdot,\cdot)$ is the second fundamental form [34] of Θ_f defined by

$$\Pi(\boldsymbol{a},\boldsymbol{b}) = -\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\left[\boldsymbol{a}^T\frac{\partial \boldsymbol{F}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{b}\right]_{k\times 1}$$

on $\mathcal{U}\times\mathcal{U}$ where $\mathcal{U} = \text{span}\{\boldsymbol{U}(\boldsymbol{\theta})\}$.

Xavier and Barroso use a polynomial bound on the cotangent to solve the lower bound of a quadratic in terms of $\text{var}(\vartheta)$. In an earlier variant of the IVLB [74], Xavier and Barroso chose $\lambda_{\boldsymbol{\theta}}^{-1} = \max_{\boldsymbol{v}\in\mathcal{U},||\boldsymbol{v}||=1} \boldsymbol{v}^T\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{v}$ and bounded $t\cot(t) \geq 1 - \frac{2}{3}t^2$, for $0 \leq t \leq T \equiv 1.35$, in (3.15) where $t = \sqrt{C}\sqrt{\text{var}(\vartheta)}$, to bound on the variance of the estimator's geodesic to the mean by

$$\text{var}(\vartheta) \geq \frac{4C + 3\lambda_{\boldsymbol{\theta}} - \sqrt{\lambda_{\boldsymbol{\theta}}(9\lambda_{\boldsymbol{\theta}} + 24C)}}{\frac{8}{3}C^2}. \tag{3.16}$$

Unfortunately, this bound was optimistic in the limit for the simple Euclidean case ($C = 0$). In a more recent paper [75], the authors improved the bound by choosing $\lambda_{\boldsymbol{\theta}} = \text{tr}(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}))^{-1}$ and an alternative lower bound for $t\cot(t)$ to obtain

$$\text{var}(\vartheta) \geq \frac{\lambda_{\boldsymbol{\theta}}C + 1 - \sqrt{2\lambda_{\boldsymbol{\theta}}C + 1}}{\frac{1}{2}C^2\lambda_{\boldsymbol{\theta}}}. \tag{3.17}$$

Although the authors omitted a proof[8], the alternative lower bound for $t\cot(t)$ appears to be $t\cot(t) \geq 1 - \frac{1}{2}t^2$ for $0 \leq t \leq T$. An immediate benefit from this

[8]Xavier and Barroso stated in [75] that the proof would "be found in the companion paper [14]", but this companion paper appears to have never been published.

improved bound is the agreement with the simple ($C = 0$) Euclidean case, i.e., $\mathrm{var}(\vartheta) \geq \lambda_{\boldsymbol{\theta}} = \mathrm{tr}(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}))^{-1}$. This consistency is entirely due to the choice of $\lambda_{\boldsymbol{\theta}}$. Even so, greater improvement of the bound, not found in the literature, is possible in the curved scenarios ($C > 0$) at least in the bound of $t\cot(t)$. Note, for $0 \leq t < \pi$

$$\begin{aligned} t\cot(t) &= 1 + \sum_{s=1}^{\infty} \frac{(-1)^s 2^{2s} B_{2s} t^{2s}}{(2s)!} = 1 - \sum_{s=1}^{\infty} 2\left(\frac{t}{\pi}\right)^{2s} \zeta(2s) \\ &\geq 1 - \sum_{s=1}^{\infty} 2\left(\frac{t}{\pi}\right)^{2s} \frac{\pi^2}{6} = 1 - \frac{\pi^2}{3} \frac{\left(\frac{t}{\pi}\right)^2}{1 - \left(\frac{t}{\pi}\right)^2} \\ &\geq 1 - \frac{b}{2} t^2 \end{aligned}$$

where B_{2s} are Bernoulli numbers, $\zeta(\cdot)$ is the Riemann zeta function, and $b = \frac{2/3}{1-\left(\frac{T}{\pi}\right)^2} \approx 0.8177$. Then

$$\mathrm{var}(\vartheta) \geq \frac{\lambda_{\boldsymbol{\theta}} C b + 1 - \sqrt{2\lambda_{\boldsymbol{\theta}} C b + 1}}{\frac{1}{2} C^2 b^2 \lambda_{\boldsymbol{\theta}}}. \tag{3.18}$$

3.1.3 Simple Extensions to the CCRB

As with unconstrained parameters, the performance of a function of parameters often may be of greater interest. Consider a continuously differentiable function $\boldsymbol{k} : \Theta_f \to \mathbb{R}^q$. Denote the Jacobian of this transfer function to be $\boldsymbol{K}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{k}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$. We have a simple extension of the classical transformation of parameters in section 2.1.1.

Corollary 3.10. If $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$, then the variance of any unbiased estimator $\boldsymbol{S}(\boldsymbol{x})$ of

$\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta})$ satisfies the inequality

$$\mathrm{Var}(\boldsymbol{S}(\boldsymbol{x})) \geq \mathrm{CCRB}(\boldsymbol{\alpha}) \triangleq \boldsymbol{K}(\boldsymbol{\theta})\mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{K}^T(\boldsymbol{\theta}).$$

Proof. Let $\boldsymbol{g}_{\boldsymbol{\theta}}$ be the implicit function defined by $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$. Then $\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}))$ has a Jacobian $\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. From the classical transformation of parameters, the inequality $\mathrm{Var}(\boldsymbol{S}(\boldsymbol{x})) \geq \mathrm{CRB}(\boldsymbol{\alpha})$ holds where

$$\begin{aligned}
\mathrm{CRB}(\boldsymbol{\alpha}) &= \boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi})\boldsymbol{G}_{\boldsymbol{\theta}}^T(\boldsymbol{\xi})\boldsymbol{K}^T(\boldsymbol{\theta}) \\
&= \boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{K}^T(\boldsymbol{\theta}).
\end{aligned}$$

□

This transformation property is useful in extending the constrained Cramér-Rao bound to biased estimation.

Example 3.11 (Biased Estimation)**.** Assume $\boldsymbol{t}(\boldsymbol{x})$ is a biased estimator of a constrained parameter $\boldsymbol{\theta}$ with bias $\boldsymbol{b}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta}$ and constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$. Define $\boldsymbol{k}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})$. Then $\boldsymbol{t}(\boldsymbol{x})$ is an unbiased estimator of $\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta})$. Since $\boldsymbol{K}(\boldsymbol{\theta}) = \boldsymbol{I}_{m\times m} + \boldsymbol{B}(\boldsymbol{\theta})$ where $\boldsymbol{B}(\boldsymbol{\theta}) = \left.\frac{\partial \boldsymbol{b}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T}\right|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$, then we have the inequality

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \left(\boldsymbol{I}_{m\times m} + \boldsymbol{B}(\boldsymbol{\theta})\right)\mathrm{CCRB}(\boldsymbol{\theta})\left(\boldsymbol{I}_{m\times m} + \boldsymbol{B}^T(\boldsymbol{\theta})\right) \triangleq \mathrm{CCRB}(\boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})).$$

Often, when the Fisher information matrix is singular, its pseudoinverse is used as a bound on the variance of an estimator. As mentioned in section 2.1.1, the bound is trivially true for some component of the estimator except under certain conditions.

Corollary 3.12. If $\boldsymbol{t}(\boldsymbol{x})$ is an estimator of $\boldsymbol{k}(\boldsymbol{\theta})$ having bias $\boldsymbol{b}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ satisfies the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$, then the inequality

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}),$$

where $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{K}(\boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{\theta})$, is nontrivially satisfied, i.e., all components of $\boldsymbol{t}(\boldsymbol{x})$ have finite variance, if and only if

$$\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}).$$

Proof. Suppose $\boldsymbol{t}(\boldsymbol{x})$ is an estimator of $\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta})$ with bias $\boldsymbol{b}(\boldsymbol{\theta})$ and under the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$. Define $\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'^T}\left(\boldsymbol{k}(\boldsymbol{\theta}) + \boldsymbol{b}(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = \boldsymbol{K}(\boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{\theta})$. If $\boldsymbol{g}_{\boldsymbol{\theta}}$ is the implicit function defined by $\boldsymbol{f}$, then we have the inequality

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \tilde{\boldsymbol{H}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^\dagger(\boldsymbol{\xi})\tilde{\boldsymbol{H}}^T(\boldsymbol{\xi})$$

where $\tilde{\boldsymbol{H}}(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}'^T}\left(\boldsymbol{k}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}')) + \boldsymbol{b}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}'))\right)\Big|_{\boldsymbol{\xi}'=\boldsymbol{\xi}} = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. (This can also be inferred from [23, lemma 2], although no assumption is made here about the singularity of the Fisher information.) Hence, all the components of $\boldsymbol{t}(\boldsymbol{x})$ can have finite variance if and only if [66]

$$\begin{aligned}
\tilde{\boldsymbol{H}}(\boldsymbol{\xi}) &= \tilde{\boldsymbol{H}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^\dagger(\boldsymbol{\xi}) \\
\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}) &= \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^\dagger(\boldsymbol{\xi})\tilde{\boldsymbol{I}}(\boldsymbol{\xi}) \\
&= \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^\dagger(\boldsymbol{\xi})\boldsymbol{G}_{\boldsymbol{\theta}}^T(\boldsymbol{\xi})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}) \\
&= \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}) \\
\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) &= \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}).
\end{aligned}$$

□

By definition $\boldsymbol{U}(\boldsymbol{\theta})$ will always be full column rank, so singularity of the matrix $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ implies singularity in the FIM $\boldsymbol{I}(\boldsymbol{\theta})$ and insufficiency in the constraints to resolve the inherent ambiguities in the model. This scenario will be discussed in greater detail in the next section.

3.1.4 Properties of the CCRB

The constrained Cramér-Rao bound for the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$ is equivalent to the CRB for any reparameterization of the parameters satisfying the constraint. Implicit in that equivalence and in the proof of the CCRB presented in section 3.1.1 are the regularity conditions for the CRB on the implicit parameter $\boldsymbol{\xi}$, i.e., $\frac{\partial}{\partial \boldsymbol{\xi}^T} E_{\boldsymbol{\xi}}(\boldsymbol{h}(x)) = E_{\boldsymbol{\xi}}(\boldsymbol{h}(x)\tilde{\boldsymbol{s}}^T(\boldsymbol{x};\boldsymbol{\xi}))$ for $\boldsymbol{h}(x) \equiv 1$ and $\boldsymbol{h}(x) \equiv \boldsymbol{t}(x)$, where $\boldsymbol{t}(x)$ in this case is an unbiased estimator of $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\xi})$. This condition translates to $\frac{\partial}{\partial \boldsymbol{\theta}^T} E_{\boldsymbol{\xi}}(\boldsymbol{h}(x))\boldsymbol{G}(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}}(\boldsymbol{h}(x)\boldsymbol{s}^T(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\xi}))$ or, strictly in terms of $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial \boldsymbol{\theta}^T} E_{\boldsymbol{\theta}}(\boldsymbol{h}(x))\boldsymbol{U}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\boldsymbol{h}(x)\boldsymbol{s}^T(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})) \tag{3.19}$$

for $\boldsymbol{h}(x) \equiv 1$ and $\boldsymbol{h}(x) \equiv \boldsymbol{t}(x)$. From this condition, it can be shown that

$$E_{\boldsymbol{\theta}}(\boldsymbol{t}(x) - \boldsymbol{\theta})\boldsymbol{s}^T(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}^T(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}^T(\boldsymbol{\theta}), \tag{3.20}$$

which is a variant of the regularity condition required and proven by Marzetta [47] and the same regularity condition simply stated (but not proven) by Stoica and Ng [68] in their proofs of the CCRB. Thus, as with the CCRB, the regularity condition for the constraint is equivalent to the regularity condition for the reparameterization of parameters satisfying the constraint. This general fact, a restatement of theorem

3.5, is the quintessential property of the CCRB formula. More explicitly, the CCRB is a generalization of both the degenerate and the determinate constraint cases, as examples 3.13 and 3.14 demonstrate.

Example 3.13 (Degenerate Case)**.** The scenario with no constraint is equivalent to the statement that the function which describes the constraint is null. That is, $\boldsymbol{f} : \Theta \to \mathbb{R}^0$ and $\boldsymbol{f}(\boldsymbol{\theta}) = [\,]$. Then $\boldsymbol{F} : \Theta \to \mathbb{R}^{0\times m}$ is also a null gradient row vector having rank 0. Any nonsingular $m \times m$ matrix $\boldsymbol{U}(\boldsymbol{\theta})$ satisfies (3.6), and thus

$$\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{U}^T(\boldsymbol{\theta}) = \boldsymbol{I}^{-1}(\boldsymbol{\theta}).$$

Therefore the CCRB formula also incorporates the unconstrained scenario.

Example 3.14 (Determinate Case)**.** Suppose the constraint $\boldsymbol{f}$ completely determines the parameter. Then necessarily, since we have m unknowns in the parameter vector $\boldsymbol{\theta}$, there must be at least $k \geq m$ equations in the constraint equation and the Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$ must have rank m. Since $\boldsymbol{f}$ is assumed to have nonredundant constraints, $\boldsymbol{F}(\boldsymbol{\theta})$ is actually a nonsingular square matrix. Only the null vector $\boldsymbol{U} : \Theta \to \mathbb{R}^{m\times 0}$ satisfies (3.6) ($\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{I}_{0\times 0}$), therefore $\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{U}^T(\boldsymbol{\theta})$ is a null element, and the CCRB does not exist for a completely determined parameter set.

Thus far, only parametric equality constraints have been considered. Gorman and Hero show that only *active* constraints results in a reduction in the bound [23, lemma 4], i.e., *inactive* (or strict) inequality constraints do not contribute information to the model in the CCRB sense.[9] As the test points approach the parameter

[9]Because the CCRB (CRB) is a local bound that only accounts for local fluctuations of the

in the Chapman-Robbins bound, they become interior points of the constraint set and the inequalities have no impact on the information metric. This can also be shown without resorting to the Chapman-Robbins approach.

Example 3.15 (Inequality Constraints Are Non-Informative)**.** Assume, in addition to the constraint set Θ_f, the parameters are required to satisfy the strict inequality $h(\boldsymbol{\theta}) < 0$ where $h : \Theta \to \mathbb{R}$ is a continuously differentiable function. To incorporate this constraint, we introduce a *dummy* parameter ϑ and add a new equality constraint $\vartheta^2 = -h(\boldsymbol{\theta})$, which is equivalent to the strict inequality constraint (whenever $\vartheta \neq 0$). In addition we create an extended parameterization $\boldsymbol{\phi}' = \begin{bmatrix} \boldsymbol{\theta}' \\ \vartheta' \end{bmatrix} \in \mathbb{R}^{m+1}$ and constraint function $\boldsymbol{f}^\star(\boldsymbol{\phi}') = \begin{bmatrix} \boldsymbol{f}(\boldsymbol{\theta}') \\ \vartheta'^2 + h(\boldsymbol{\theta}') \end{bmatrix}$. This will generate a Fisher information and a Jacobian matrix defined by

$$\boldsymbol{I}^\star(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{I}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad , \quad \boldsymbol{F}^\star(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{F}(\boldsymbol{\theta}) & \mathbf{0} \\ \boldsymbol{H}^T(\boldsymbol{\theta}) & 2\vartheta \end{bmatrix},$$

respectively, where $\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial h(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$. Note that $\boldsymbol{F}^\star(\boldsymbol{\phi})$ will be full row rank as long as $\vartheta \neq 0$. If $\boldsymbol{U}(\boldsymbol{\theta})$ is defined as in (3.6), then $\boldsymbol{U}^\star(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\theta}) \\ \boldsymbol{v}^T(\boldsymbol{\theta}, \vartheta) \end{bmatrix}$ will also satisfy (3.6) with respect to $\boldsymbol{F}^\star(\boldsymbol{\phi})$, where $\boldsymbol{v}(\boldsymbol{\theta}, \vartheta) = \frac{1}{2\vartheta}\boldsymbol{H}^T(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$. By theorem

true parameter, inactive constraints have no impact on performance potential. As such the CCRB (CRB) only provides information on the pdf for the mainlobe of the density function, which typically corresponds to the true parameter. For a number of scenarios, e.g., when there is sufficiently large variance in the pdf, sidelobes of the distribution impact the performance, thereby making the CCRB (CRB) overly optimistic. This occurs frequently in communications when the signal-to-noise ratio (SNR) or data transmission size decreases.

3.1, then

$$\begin{aligned}
&\boldsymbol{U}^{\star}(\boldsymbol{\phi})\left(\boldsymbol{U}^{\star T}(\boldsymbol{\phi})\boldsymbol{I}^{\star}(\boldsymbol{\phi})\boldsymbol{U}^{\star}(\boldsymbol{\phi})\right)^{-1}\boldsymbol{U}^{\star T}(\boldsymbol{\phi}) \\
&\quad = \boldsymbol{U}^{\star}(\boldsymbol{\phi})\left(\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^{\star T}(\boldsymbol{\phi}) \\
&\quad = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}(\boldsymbol{\theta}) & \boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{v}(\boldsymbol{\theta},\vartheta) \\ \boldsymbol{v}^{T}(\boldsymbol{\theta},\vartheta)\left(\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^{T}(\boldsymbol{\theta}) & \boldsymbol{v}^{T}(\boldsymbol{\theta},\vartheta)\left(\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{v}(\boldsymbol{\theta},\vartheta) \end{bmatrix}.
\end{aligned}$$

Hence, the CCRB on the $\boldsymbol{\theta}$ components of $\boldsymbol{\phi}$, the upper-left submatrix, remains unchanged.

Equality constraints, however, do add side information to the model. Thus, it is intuitive to expect that the constrained model should result in a lower bound compared to the bound for the model without constraints. This statement was made in Gorman and Hero [23, p.1292], but not in Marzetta [47] nor Stoica and Ng [68]. In the latter case, the statement is only true under certain conditions. Prior to establishing when the bound is lowered, a powerful lemma will be proven.

Lemma 3.16. For an arbitrary full column rank matrix $\boldsymbol{A}$, and an arbitrary symmetric positive semidefinite matrix $\boldsymbol{B}$, the inequality

$$\boldsymbol{A}\left(\boldsymbol{A}^{T}\boldsymbol{B}\boldsymbol{A}\right)^{\dagger}\boldsymbol{A}^{T} \leq \boldsymbol{B}^{\dagger} \tag{3.21}$$

holds over the projection subspace of $\boldsymbol{B}^{\dagger}\boldsymbol{B}$ with equality if and only if $\operatorname{rank}(\boldsymbol{A}^{T}\boldsymbol{B}\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{B})$.

Proof. Let $\boldsymbol{L}\boldsymbol{L}^{T}$ be the Cholesky decomposition of $\boldsymbol{B}$ [20, p. 194]. Then $\boldsymbol{L} \in \mathbb{R}^{m\times\operatorname{rank}(\boldsymbol{B})}$ is a full column rank lower triangular matrix with strictly positive values on the diagonal. To show the inequality, consider linear unbiased estimates of the

mean in the model $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{L}^T\boldsymbol{L})^{-2})$, and $\boldsymbol{\beta}$ is treated as the unknown parameter. In particular, $\boldsymbol{y}$ is such an estimate with variance equaling $(\boldsymbol{L}^T\boldsymbol{L})^{-2}$ where as the best linear unbiased estimate (see section 2.3.1)

$$\boldsymbol{D}\hat{\boldsymbol{\beta}} = \boldsymbol{D}\left(\boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{D}\right)^\dagger \boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{y},$$

with variance equal to $\boldsymbol{D}\left(\boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{D}\right)^\dagger \boldsymbol{D}^T$. By the Gauss-Markov theorem we have the inequality with the BLUE's variance

$$\boldsymbol{D}\left(\boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{D}\right)^\dagger \boldsymbol{D}^T \leq (\boldsymbol{L}^T\boldsymbol{L})^{-2}$$

with equality if and only if $(\boldsymbol{L}^T\boldsymbol{L})\boldsymbol{D}\left(\boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{D}\right)^\dagger \boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L}) = \boldsymbol{I}_{\mathrm{rank}(\boldsymbol{L})\times\mathrm{rank}(\boldsymbol{L})}$. Substituting in $\boldsymbol{D} = \left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T\boldsymbol{A}$ in 3.1.4 and pre- and post-multiplying both sides by $\boldsymbol{L}$ and $\boldsymbol{L}^T$, respectively, we have the inequality

$$\boldsymbol{B}^\dagger\boldsymbol{B}\boldsymbol{A}\left(\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{A}\right)^\dagger \boldsymbol{A}^T\boldsymbol{B}\boldsymbol{B}^\dagger \leq \boldsymbol{B}^\dagger.$$

Considering quadratic forms in $\boldsymbol{B}^\dagger\boldsymbol{B}\boldsymbol{v}$ proves the result since for any vector $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{v}^T\boldsymbol{B}^\dagger\boldsymbol{B}\boldsymbol{B}^\dagger\boldsymbol{B}\boldsymbol{B}^\dagger\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{B}^\dagger\boldsymbol{v}$. Moreover, from the definition of $\boldsymbol{D}$, we have $(\boldsymbol{L}^T\boldsymbol{L})\boldsymbol{D}\left(\boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L})^2\boldsymbol{D}\right)^\dagger \boldsymbol{D}^T(\boldsymbol{L}^T\boldsymbol{L}) = \boldsymbol{I}_{\mathrm{rank}(\boldsymbol{L})\times\mathrm{rank}(\boldsymbol{L})}$ if and only if it can be shown that $\boldsymbol{L}^T\boldsymbol{A}\left(\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{A}\right)^\dagger \boldsymbol{A}^T\boldsymbol{L} = \boldsymbol{I}_{\mathrm{rank}(\boldsymbol{L})\times\mathrm{rank}(\boldsymbol{L})}$, which is so if and only if $\boldsymbol{A}$ and $\boldsymbol{L}$ satisfy $\mathrm{rank}(\boldsymbol{A}^T\boldsymbol{L}) = \mathrm{rank}(\boldsymbol{B})$. □

This lemma proves the following results, including that in a linear subspace the CCRB($\boldsymbol{\theta}$) is lesser than or equal to the CRB($\boldsymbol{\theta}$) in the matrix sense and that constraints strictly on the ambiguous information have no impact on the performance of the unambiguous information.

Theorem 3.17. Let $\boldsymbol{U}(\boldsymbol{\theta})$ be defined as in (3.6) from a constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$. Then

$$\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta}) \leq \boldsymbol{I}^\dagger(\boldsymbol{\theta})$$

over the linear projection subspace defined by $\boldsymbol{I}^\dagger(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})$.

Proof. From the lemma, defining $\boldsymbol{A} = \boldsymbol{U}(\boldsymbol{\theta})$ and $\boldsymbol{B} = \boldsymbol{I}(\boldsymbol{\theta})$ gives the inequality. □

The conditions under which $\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^\dagger \boldsymbol{U}^T(\boldsymbol{\theta})$ and $\boldsymbol{I}^\dagger(\boldsymbol{\theta})$ are nontrivial bounds are detailed in corollary 3.12 and section 2.1.1, respectively. The projection space of $\boldsymbol{I}^\dagger(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})$ corresponds to the identifiable components of $\boldsymbol{\theta}$ without the constraints (e.g., see [18, section 4]). The theorem establishes the result that the constraints can only lower the bound and thereby increase performance potential for (functions of) parameters that are already identifiable. This is regardless of whether the Fisher information is singular or whether the parameters are identifiable under constraints (whether $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is singular, see Theorem 3.24). Theorem 3.17 is more general than the result shown by Ash[10], which is given by the following corollary.

Corollary 3.18 (Ash)**.** Provided no linear combination of $\boldsymbol{U}(\boldsymbol{\theta})$ lies in the null space of the Fisher information matrix, then

$$\boldsymbol{U}(\boldsymbol{\theta}) \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{U}^T(\boldsymbol{\theta}) \leq \boldsymbol{I}^\dagger(\boldsymbol{\theta})$$

over the linear projection subspace defined by $\boldsymbol{I}^\dagger(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})$.

[10] Although not stated in this manner, this result appears in Ash's thesis [5, equation (3.63)] as well as in a publication of his third chapter in Ash and Moses [6, equation (63)]. The results from Lemma 3.16 are more general than those in [5] or [6].

Proof. Since $\boldsymbol{U}(\boldsymbol{\theta})$ is full column rank and the range space of its columns does not exist in the null space of the FIM, then $\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)$ is nonsingular. □

Furthermore, if the Fisher information matrix is nonsingular, then $\boldsymbol{I}^{\dagger}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})$ is the identity matrix, so the inequality holds as a quadratic form over $\mathbb{R}^m$ and

$$\boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta}) \leq \boldsymbol{I}^{-1}(\boldsymbol{\theta}).$$

This was also cleverly shown by Gorman and Hero [23, (44) in remark 4], again, with the unnecessary assumption that $\boldsymbol{I}(\boldsymbol{\theta})$ is regular (non-singular). Next, we use the lemma to observe the existence of *non-informative* constraints.

Corollary 3.19. Assume the row vectors of $\boldsymbol{F}(\boldsymbol{\theta}) \in \mathbb{R}^{k\times m}$ form a linearly independent basis for the null space of $\boldsymbol{I}(\boldsymbol{\theta})$. Then $\boldsymbol{F}_1(\boldsymbol{\theta}) \in \mathbb{R}^{k'\times m}$ is a linear combination of a submatrix of $\boldsymbol{F}(\boldsymbol{\theta})$ if and only if

$$\boldsymbol{U}_1(\boldsymbol{\theta})\left(\boldsymbol{U}_1^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}_1(\boldsymbol{\theta})\right)^{\dagger}\boldsymbol{U}_1^T(\boldsymbol{\theta}) = \boldsymbol{I}^{\dagger}(\boldsymbol{\theta}),$$

where $\boldsymbol{U}_1(\boldsymbol{\theta})$ is defined as in (3.6) relative to $\boldsymbol{F}_1(\boldsymbol{\theta})$.

Proof. Without loss of generality, partition $\boldsymbol{F}$ as $\begin{bmatrix}\boldsymbol{F}_1(\boldsymbol{\theta})\\ \boldsymbol{F}_2(\boldsymbol{\theta})\end{bmatrix}$. Then if $\boldsymbol{U}(\boldsymbol{\theta})$ is defined as in (3.6), define

$$\boldsymbol{U}_1(\boldsymbol{\theta}) = \left[\boldsymbol{U}(\boldsymbol{\theta}),\quad \left(I_{m\times m} - \boldsymbol{F}_1^T\left(\boldsymbol{F}_1\boldsymbol{F}_1^T\right)^{-1}\boldsymbol{F}_1\right)\boldsymbol{F}_2^T(\boldsymbol{\theta})\boldsymbol{D}(\boldsymbol{\theta})\right],$$

where $\boldsymbol{D}(\boldsymbol{\theta})$ represents a Gram-Schmidt processing matrix which orthonormalizes the column vectors of $\left(I_{m\times m} - \boldsymbol{F}_1^T\left(\boldsymbol{F}_1\boldsymbol{F}_1^T\right)^{-1}\boldsymbol{F}_1\right)\boldsymbol{F}_2^T(\boldsymbol{\theta})$ $\boldsymbol{F}_2^T(\boldsymbol{\theta})$ (these are already orthogonal to the column vectors of $\boldsymbol{U}(\boldsymbol{\theta})$). This satisfies (3.6). Since $\operatorname{span}(\boldsymbol{I}(\boldsymbol{\theta})) \subset$

span($\boldsymbol{U}_1(\boldsymbol{\theta})$) then rank($\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})$) = rank($\boldsymbol{I}(\boldsymbol{\theta})$), and thus, by lemma 3.16,

$$\boldsymbol{U}_1(\boldsymbol{\theta})\left(\boldsymbol{U}_1^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}_1(\boldsymbol{\theta})\right)^{\dagger}\boldsymbol{U}_1^T(\boldsymbol{\theta}) = \boldsymbol{I}^{\dagger}(\boldsymbol{\theta}).$$

□

This corollary details the necessary and sufficient characteristics for a constraint to be non-informative. Implicit from the lemma and above corollary is the CCRB's invariance to linear combinations of columns of $\boldsymbol{U}(\boldsymbol{\theta})$ that exist in the null space of $\boldsymbol{I}(\boldsymbol{\theta})$. The specific case when span($\boldsymbol{I}(\boldsymbol{\theta})$) = span($\boldsymbol{U}(\boldsymbol{\theta})$) is also shown by Ash in [5, equation (3.73)] and [6, equation (73)].[11]

The CCRB can be interpreted geometrically, as in Figure 3.1 as a contraction of the information ambiguity to find the bound and then an expansion. The column vectors of $\boldsymbol{U}(\boldsymbol{\theta})$, being in the null space of the Jacobian of the constraints, restrict the information in Θ into the constraint space Θ_f. The bound can be found from (the inverse of) the information in Θ_f, and is then projected back into the original space of the parameters of interest. This down-and-up projection is easiest to observe for linear constraints, where any local reparameterization is also a global reparameterization.

Example 3.20. Assume the parameters satisfy the linear constraint

$$\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{F}\boldsymbol{\theta} + \boldsymbol{v} = \boldsymbol{0}.$$

In this case, solutions of $\boldsymbol{\theta}$ are of the form $\boldsymbol{\theta} = -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} + \boldsymbol{U}\boldsymbol{\xi} = \boldsymbol{g}(\boldsymbol{\xi})$ where

[11] In Ash's thesis and paper, this scenario was referred to as the *minimally constrained* case, where only all the unknown information ambiguities are constrained. However, in this current work, the minimal number of constraints is zero, the degenerate constraint case of example 3.13, thus it makes more sense to refer to this scenario as a non-informative constraint.

$\boldsymbol{U}$ is defined as in (3.6). That this is a one-to-one correspondence locally (or in this case globally) is made clear by the existence of the inverse $\boldsymbol{g}^{-1}(\boldsymbol{\theta}) = \boldsymbol{U}^T\boldsymbol{\theta} = \boldsymbol{\xi}$. Then, since $\boldsymbol{G}(\boldsymbol{\xi}) = \boldsymbol{U}$, we have from (3.9)

$$\text{CCRB}(\boldsymbol{\theta}) = \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T$$

which is exactly the CCRB provided in (3.5).

Alternatively, another interpretation [9] is that the CCRB is less than (in a matrix sense) the CRB because the performance bound is over an expanded class of estimators. The (non-biased) CRB is a bound on the mean-square error of unbiased estimators in Θ, whereas the CCRB is a bound for estimators that only need to be unbiased on Θ_f and not on the whole set Θ.

3.1.5 Derivation of $\boldsymbol{g}(\boldsymbol{\xi})$

In the proof of the CCRB as well as its application, the need for an explicit reparameterization $\boldsymbol{g}_{\boldsymbol{\theta}}$ proved unnecessary. However, in examples 3.4 and 3.20 of the previous section, we presented scenarios where, given the constraint function $\boldsymbol{f}$, we were able to define a locally equivalent continuously differentiable $\boldsymbol{g}_{\boldsymbol{\theta}}$. There may exist other scenarios where it is desirable to obtain a $\boldsymbol{g}_{\boldsymbol{\theta}}$ explicitly.

In this section, we present two procedures to do so. First, we detail an approach using the Taylor expansion of $\boldsymbol{g}_{\boldsymbol{\theta}}$ given explicit knowledge of $\boldsymbol{U}(\boldsymbol{\theta}')$. Then, we demonstrate an approach based on fixed point methods.

3.1.5.1 A Taylor series derivation

The Taylor series expansion of $\boldsymbol{\theta}' = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}')$ about $\boldsymbol{\xi}$ is given by

$$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}') = \boldsymbol{\theta} + \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})(\boldsymbol{\xi}' - \boldsymbol{\xi}) + \cdots$$

where $\boldsymbol{\theta} = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. For any differentiable function $h : \mathbb{R}^m \to \mathbb{R}$ we have

$$\frac{\partial}{\partial \boldsymbol{\xi}'^T} h(\boldsymbol{\theta}') = \frac{\partial}{\partial \boldsymbol{\theta}'^T} h(\boldsymbol{\theta}') \cdot \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}').$$

Since the implicit function from $\mathbb{R}^{m-k}$ into Θ_f is not unique, we can choose a reparameterization which uses the transformation matrix $\boldsymbol{S}(\boldsymbol{\theta}) = \boldsymbol{I}_{m-k}$, i.e., we choose a reparameterization with Jacobian $\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}') = \boldsymbol{U}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}'))$ for any null space matrix $\boldsymbol{U}(\boldsymbol{\theta})$ that satisfies (3.6) and choose $\boldsymbol{\xi} = \boldsymbol{0}$. With this selection of the reparameterization, the rth order derivatives of $\boldsymbol{g}_{\boldsymbol{\theta}}$ are the $(r-1)$st order derivatives of the elements of $\boldsymbol{U}(\boldsymbol{\theta}')$ with respect to $\boldsymbol{\xi}'$, which are to be evaluated at $\boldsymbol{\theta}$ for the coefficients in the Taylor series.

Example 3.21. Reviewing example 3.4, note

$$\boldsymbol{U}(\boldsymbol{\theta}') = \begin{bmatrix} \theta_2' \\ -\theta_1' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1' \\ \theta_2' \end{bmatrix}$$

and hence

$$\frac{\partial \boldsymbol{U}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{3.22}$$

is independent of $\boldsymbol{\theta}'$, so using (3.22) as a reference, we have

$$\frac{\partial^r \boldsymbol{U}(\boldsymbol{\theta}')}{\partial \boldsymbol{\xi}'^r} = \frac{\partial^{r-1}}{\partial \boldsymbol{\xi}'^{r-1}} \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T} \cdot \boldsymbol{U}(\boldsymbol{\theta}') \right) = \cdots = \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T} \right)^r \boldsymbol{U}(\boldsymbol{\theta}').$$

With a natural selection of $\boldsymbol{\theta} = \boldsymbol{g}_{\boldsymbol{\theta}}(0) = (1,0)^T$ as an initial value for the Taylor series, the reparameterized function is found to be

$$\begin{aligned}
\boldsymbol{\theta}' = \boldsymbol{g}(\xi') &= \sum_{r=0}^{\infty} \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'^T} \right)^r \boldsymbol{\theta} \frac{\xi'^r}{r!} \\
&= \sum_{r=0}^{\infty} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}^r \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{\xi'^r}{r!} \\
&= \begin{bmatrix} \sum_{r=0}^{\infty} (-1)^r \frac{\xi'^{2r}}{(2r)!} \\ \sum_{r=0}^{\infty} (-1)^{r+1} \frac{\xi'^{2r+1}}{(2r+1)!} \end{bmatrix} \\
&= \begin{bmatrix} \cos(\xi') \\ -\sin(\xi') \end{bmatrix}.
\end{aligned}$$

This particular choice of $\boldsymbol{\theta}$ produces a reparameterization $\boldsymbol{g}_{\boldsymbol{\theta}}$ in agreement with the one chosen in (3.8). Any alternative reparameterization can be found utilizing an alternative transformation matrix $\boldsymbol{S}(\boldsymbol{\theta}')$ or an alternative initialization.

Such an approach will only derive a local bijective map and convergence of the Taylor series may not result in a known functional form for any given constraint $\boldsymbol{f}$. When $\boldsymbol{U}(\boldsymbol{\theta}')$ is not known as a function of $\boldsymbol{\theta}'$, numerical techniques are available to find the derivatives of $\boldsymbol{U}(\boldsymbol{\theta}')$ with respect to $\boldsymbol{\xi}'$ using the equation $\boldsymbol{F}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}') = \boldsymbol{0}$.

3.1.5.2 A fixed point derivation

More commonly, a fixed point approach is taken to the derivation of an implicit function. As in appendix A.2, the parameter vector is partitioned $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}$ and the constraint function rewritten as $\boldsymbol{f} : \mathbb{R}^{m-k} \times \mathbb{R}^k \rightarrow \mathbb{R}^k$ defined as $\boldsymbol{f}^\star(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2) = \boldsymbol{f}(\begin{bmatrix} \boldsymbol{\theta}'_1 \\ \boldsymbol{\theta}'_2 \end{bmatrix})$. If $\boldsymbol{f}^\star_{\boldsymbol{\theta}'_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{\partial}{\partial \boldsymbol{\theta}'^T_2} \boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_2)\Big|_{\boldsymbol{\theta}'_2 = \boldsymbol{\theta}_2}$ is nonsingular, then there exist a

unique continuous function $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)$ about $\boldsymbol{\theta}_1$ such that $\boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)) = \mathbf{0}$. This is merely a particular variation of the implicit function theorem. One proof of this version proves the existence of a fixed point in the contraction map

$$\boldsymbol{d}(\boldsymbol{\theta}_2(\boldsymbol{\theta}_1')) = \boldsymbol{\theta}_2(\boldsymbol{\theta}_1') - \boldsymbol{D}^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{f}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2(\boldsymbol{\theta}_1'))$$

where $\boldsymbol{D}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{f}^\star_{\boldsymbol{\theta}_2'}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. This motivates the use of the iteration

$$\boldsymbol{\theta}_2^{(i+1)}(\boldsymbol{\theta}_1') = \boldsymbol{\theta}_2^{(i)}(\boldsymbol{\theta}_1') - \boldsymbol{D}^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{f}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2^{(i)}(\boldsymbol{\theta}_1'))$$

to generate the (fixed point) implicit function. The iteration is essentially an application of Newton's method [64].

Example 3.22. Reviewing example 3.4 again, we have the constraint $f^\star(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 - 1 = 0$ and wish to find a local reparameterization at $(\theta_1, \theta_2) = (1, 0)$. First note,

$$f^\star_{\theta_1}(\theta_1, \theta_2) = 2\theta_1 = 2 \quad , \quad f^\star_{\theta_2}(\theta_1, \theta_2) = 2\theta_2 = 0,$$

and we only have a nonsingularity with respect to θ_1. Hence, we cannot find a find a function for θ_2 in terms of θ_1 using this approach at $(1, 0)$, but we can find $\theta_1'(\theta_2')$ there. Defining $D(\theta_1, \theta_2) = f^\star_{\theta_1}(\theta_1, \theta_2) = 2$ and initializing with $\theta_1^{'(1)}(\theta_2') = 1$, the fixed point method dictates the next iterate to be

$$\begin{aligned} \theta_1^{'(2)}(\theta_2') &= \theta_1^{'(1)} - D^{-1}(\theta_1, \theta_2)\left((\theta_1^{'(1)}(\theta_2'))^2 + \theta_2'^2 - 1\right) \\ &= 1 - \frac{1}{2}\left(1 + \theta_2'^2 - 1\right) = 1 - \frac{1}{2}\theta_2'^2. \end{aligned}$$

Continuing, the third iterate is

$$\begin{aligned} \theta_1^{'(3)}(\theta_2') &= 1 - \frac{1}{2}\theta_2'^2 - \frac{1}{2}\left(\left(1 - \tfrac{1}{2}\theta_2'^2\right)^2 + \theta_2'^2 - 1\right) \\ &= 1 - \frac{1}{2}\theta_2'^2 - \frac{1}{8}\theta_2'^4, \end{aligned}$$

and the fourth iterate is

$$\begin{aligned}\theta_1^{'(4)}(\theta_2') &= 1 - \frac{1}{2}\theta_2'^2 - \frac{1}{8}\theta_2'^4 - \frac{1}{2}\left(\left(1 - \tfrac{1}{2}\theta_2'^2 - \tfrac{1}{8}\theta_2'^4\right)^2 + \theta_2'^2 - 1\right) \\ &= 1 - \frac{1}{2}\theta_2'^2 - \frac{1}{8}\theta_2'^4 - \frac{1}{16}\theta_2'^6 - \frac{1}{128}\theta_2'^8.\end{aligned}$$

As expected, as $r \to \infty$ then $\theta_1^{'(r)}(\theta_2')$ approaches a limit function

$$\theta_1'(\theta_2') = \sqrt{1 - \theta_2'^2} = 1 - \frac{1}{2}\theta_2'^2 - \frac{1}{8}\theta_2'^4 - \frac{1}{16}\theta_2'^6 - \frac{5}{128}\theta_2'^8 + O(\theta_2'^9)$$

near $\theta_2 = 0$.

3.2 Identifiability

Identifiability conditions based on the CRB (or FIM) were detailed in Section 2.2, and the definitions of local and strong identifiability are given therein. In this section, the identifiability of parameters under functional equality constraints is considered.

3.2.1 Local identifiability

To establish a new identifiability criterion from the CCRB, we will first examine an existing criterion. Rothenberg [58, theorem 6] developed conditions for identifiability of a parameter vector $\boldsymbol{\theta}$ under the constraints $\boldsymbol{f}(\boldsymbol{\theta}) = 0$, which was later partially re-derived by Crowder [18, lemma 1 is the "only if" portion of the theorem statement].

Theorem 3.23 (Rothenberg-Crowder)**.** Assume both $\boldsymbol{F}(\boldsymbol{\theta}')$ and

$$\boldsymbol{M}(\boldsymbol{\theta}') = \begin{bmatrix}\boldsymbol{I}(\boldsymbol{\theta}') \\ \boldsymbol{F}(\boldsymbol{\theta}')\end{bmatrix}$$

have constant rank in a local neighborhood about $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}$ is locally identifiable if and only if $\boldsymbol{M}(\boldsymbol{\theta})$ has full column rank m.

The proof of this theorem is based on the unconstrained proof of Theorem 2.5. An immediate implication is that if the FIM $\boldsymbol{I}(\boldsymbol{\theta})$ is regular, then not only is the unconstrained model locally identifiable, but the constrained model is as well. Otherwise, if the FIM is singular, then the constraints must be such that the row vectors of its Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$ eliminate the null space of the FIM. In doing so, the constraints eliminate whatever inherent ambiguity was in the model that led to local unidentifiability and a singular FIM (Theorem 2.5). If row vectors of the Jacobian did not eliminate the null space of the Fisher information then there would exist a linear combination of column vectors of $\boldsymbol{U}(\boldsymbol{\theta})$ such that $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is singular, as shall be shown shortly.

Additionally, regardless of the information, in the trivial case where the constraints are such that the Jacobian is full column rank, then the theorem says the model is locally identifiable. Indeed, when $\text{rank}(\boldsymbol{F}(\boldsymbol{\theta})) = m$ the constraints completely determine the parameter (e.g., see example 3.14).

Rothenberg's theorem for unconstrained identifiability is useful in establishing an alternative criterion for identifiability in relation to a component of the CCRB.

Theorem 3.24. Let $\boldsymbol{\theta} \in \Theta_f$ and assume $\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')$ has constant rank in a neighborhood of $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}$ is locally identifiable if and only if $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is regular.

Proof. Let $\boldsymbol{g}_{\boldsymbol{\theta}}$ satisfy Theorem 3.3. If $\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')$ has constant rank in a

local neighborhood about $\boldsymbol{\theta} = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. Then $\tilde{\boldsymbol{I}}(\boldsymbol{\xi}^{'})$ has constant rank in a local neighborhood about $\boldsymbol{\xi}$. And since $\boldsymbol{g}_{\boldsymbol{\theta}}$ is injective, $\boldsymbol{\theta}$ is locally identifiable in Θ_f if and only if $\boldsymbol{\xi}$ is locally identifiable in R^{m-k}. By Theorem 2.5, $\boldsymbol{\xi}$ is locally identifiable if and only if $\tilde{\boldsymbol{I}}(\boldsymbol{\xi})$ is regular. And $\tilde{\boldsymbol{I}}(\boldsymbol{\xi})$ is regular if and only if $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is regular. □

This is the corresponding theorem to Rothenberg's Theorem 2.5 and agrees with his Theorem 3.23, although the proof does not rely on this latter theorem's result because the implicit function $\boldsymbol{g}_{\boldsymbol{\theta}}$ simplifies the approach. It is, however, possible to prove a more general result connecting the rank of the $\boldsymbol{M}(\cdot)$ matrix of Theorem 3.23 to the implicit Fisher information $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$.

Theorem 3.25. Assume $\boldsymbol{U}^T(\boldsymbol{\theta}^{'})\boldsymbol{I}(\boldsymbol{\theta}^{'})\boldsymbol{U}(\boldsymbol{\theta}^{'})$ has constant rank in a neighborhood of $\boldsymbol{\theta}$. Then

$$\text{nullity}(\boldsymbol{M}(\boldsymbol{\theta}^{'})) = \text{nullity}(\boldsymbol{U}^T(\boldsymbol{\theta}^{'})\boldsymbol{I}(\boldsymbol{\theta}^{'})\boldsymbol{U}(\boldsymbol{\theta}^{'})).$$

Proof. First, for some fixed $\boldsymbol{\theta}^{'}$, assume $\boldsymbol{M}(\boldsymbol{\theta}^{'})$ is not full column rank and the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ are linearly independent and span the null space of $\boldsymbol{M}(\boldsymbol{\theta}^{'})$, i.e., for $l = 1, \ldots, r$,

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta}^{'})\boldsymbol{v}_l &= \boldsymbol{0} \\ \boldsymbol{F}(\boldsymbol{\theta}^{'})\boldsymbol{v}_l &= \boldsymbol{0}. \end{aligned}$$

Let $\boldsymbol{U}(\boldsymbol{\theta}^{'})$ be a matrix defined as in (3.6). Since $\boldsymbol{F}(\boldsymbol{\theta}^{'})\boldsymbol{v}_l = \boldsymbol{0}$ then each $\boldsymbol{v}_l = \boldsymbol{U}(\boldsymbol{\theta}^{'})\boldsymbol{w}_l$ for some $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r \in \mathbb{R}^{m-k}$. Now if $\sum_{l=1}^{r} \gamma_l \boldsymbol{w}_l = \boldsymbol{0}$, then $\sum_{l=1}^{r} \gamma_l \boldsymbol{v}_l = \sum_{l=1}^{r} \boldsymbol{U}(\boldsymbol{\theta}^{'})\gamma_l \boldsymbol{w}_l = \boldsymbol{0}$, which implies $\gamma_1 = \cdots = \gamma_r = 0$ since the $\boldsymbol{v}_i$ are linearly independent. Hence,

the $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$ are also linearly independent and span an r-dimensional subspace of $\mathbb{R}^{m-k}$. Thus, $\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')\boldsymbol{w}_l = \boldsymbol{0}$ for $l = 1, \ldots, r$. This proves the implication, namely, that nullity$(\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')) \geq$ nullity$(\boldsymbol{M}(\boldsymbol{\theta}'))$.

To show the inverse, assume the vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$ are linearly independent and span the null space of the $\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')$. Define $\boldsymbol{v}_l = \boldsymbol{U}(\boldsymbol{\theta}')\boldsymbol{w}_l$ for $l = 1, \ldots, r$. Since $\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{v}_l = \boldsymbol{0}$ and $\boldsymbol{F}(\boldsymbol{\theta}')\boldsymbol{v}_l = \boldsymbol{0}$, then $\boldsymbol{M}(\boldsymbol{\theta}')\boldsymbol{v}_1 = \boldsymbol{0}$ for $l = 1, \ldots, r$. Note for any $\gamma_1, \ldots, \gamma_r$, then $\sum_{l=1}^{r} \gamma_l \boldsymbol{v}_l = \sum_{l=1}^{r} \boldsymbol{U}(\boldsymbol{\theta}')\gamma_l \boldsymbol{w}_l = \boldsymbol{0}$ if and only if $\sum_{l=1}^{r} \gamma_l \boldsymbol{w}_l = \boldsymbol{0}$ (since $\boldsymbol{U}(\boldsymbol{\theta}')$ is full column rank) which is true if and only if $\gamma_l = \cdots = \gamma_r = 0$. This proves the converse, i.e., nullity$(\boldsymbol{M}(\boldsymbol{\theta}')) \geq$ nullity$(\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}'))$. □

The theorem essentially states that in the local neighborhood where the implicit function $\boldsymbol{g}$ is defined, $\boldsymbol{M}(\boldsymbol{\theta})$ has full column rank if and only if $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ does, for any $\boldsymbol{\theta} \in \Theta_f$. As a consequence, we have the following corollary.

Corollary 3.26. If $\boldsymbol{I}(\boldsymbol{\theta})$ is regular, then $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is also regular. And if $\boldsymbol{\theta}$ is locally identifiable in Θ, then $\boldsymbol{\theta}$ is locally identifiable in Θ_f.

Proof. If $\boldsymbol{\theta}$ is locally identifiable, then nullity$(\boldsymbol{I}(\boldsymbol{\theta})) = 0$ by theorem 2.5. And if nullity$(\boldsymbol{I}(\boldsymbol{\theta})) = 0$, then nullity$(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})) = 0$ by theorem 3.25. Finally, if nullity$(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})) = 0$, then $\boldsymbol{\theta}$ is locally identifiable in Θ_f by theorem 3.24.

□

3.2.1.1 Local identifiability in the Aitchison-Silvey-Crowder CCRB formula

In subsection 3.1.2.2, an alternative form of the CCRB is presented where a loaded Fisher information is used in place of the FIM to resolve issues of singularity of the FIM. The Aitchison-Silvey loaded FIM is $\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$ [4], whereas the Crowder loaded FIM $\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{K}\boldsymbol{F}(\boldsymbol{\theta})$ [18], where $\boldsymbol{K}$ is chosen such that the loaded FIM is full rank. The following theorem, which connects local identifiability and the Aitchison-Silvey-Crowder CCRB, shows that under certain conditions, the matrix $\boldsymbol{K}$ is unnecessary. This result is hinted at, but not clearly stated in [18, lemma 6].

Theorem 3.27. The Aitchison-Silvey-Crowder loaded FIM $\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$ is nonsingular if and only if $\boldsymbol{M}(\boldsymbol{\theta})$ is full column rank. Hence, if $\boldsymbol{I}(\boldsymbol{\theta}') + \boldsymbol{F}^T(\boldsymbol{\theta}')\boldsymbol{F}(\boldsymbol{\theta}')$ is constant locally about $\boldsymbol{\theta}$, then $\boldsymbol{\theta}$ is identifiable if and only if $\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$ is nonsingular.

Proof. To show the contrapositive, assume $\boldsymbol{M}(\boldsymbol{\theta})$ is not full column rank and $\boldsymbol{v}$ is a nontrivial vector such that $\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$. Then $\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$. Therefore, $\boldsymbol{v}$ is in the null space of the Gram matrix $\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$. To show the inverse, assume $\left(\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})\right)\boldsymbol{v} = \boldsymbol{0}$ for some nontrivial vector $\boldsymbol{v} \in \mathbb{R}^m$. Then since

$$\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{I}^{1/2}(\boldsymbol{\theta}) & \boldsymbol{F}^T(\boldsymbol{\theta})\end{bmatrix} \cdot \begin{bmatrix}\boldsymbol{I}^{1/2}(\boldsymbol{\theta}) \\ \boldsymbol{F}(\boldsymbol{\theta})\end{bmatrix},$$

we must have that $\boldsymbol{I}^{1/2}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$. Hence, $\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{I}^{1/2}(\boldsymbol{\theta})\boldsymbol{I}^{1/2}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$, which implies $\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{M}(\boldsymbol{\theta})$ is not full column rank. □

3.2.2 Strong Identifiability

Restricting the discussion to normal distributions in this section we can extend the equivalence criterion between regularity and strong identifiability as defined in section 2.2.2. That is, assume $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ with $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathbb{R}^p$, with the elements of the mean and variance explicitly defined by a map $\boldsymbol{\varphi} : \Theta \to \mathbb{R}^q$ where $q \leq p + p(p+1)/2$ and assume $m \leq q$.

Theorem 3.28. If $\boldsymbol{\theta}$ is strongly identifiable on Θ and $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, then $\boldsymbol{\theta}$ is also strongly identifiable on Θ_f.

Proof. If $\boldsymbol{\theta}$ is strongly identifiable, then there exists a representative mapping $\boldsymbol{\varphi}^*$, which is injective on Θ. Since $\Theta_f \subset \Theta$, $\boldsymbol{\varphi}^*$ is still injective and a representative mapping on Θ_f. □

This theorem is complementary to Corollary 3.26. Essentially, the imposition of constraints does not take away existing identifiability (local or strong) or Fisher information regularity that already exists in a model. However, it is not always the case that the original (unconstrained) model is information regular or identifiable. The following theorem, an extension of Theorem 2.6, connects the notion of strong identifiability with regularity of $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$.

Theorem 3.29. Assume $\boldsymbol{\varphi}$ is a holomorphic mapping of $\boldsymbol{z} \in \Omega \subset \cup_{\alpha \in A} \Omega_\alpha$ into $\mathbb{C}^q$, where $\Theta_f \subset \Omega \subset \mathbb{C}^m$ and Ω_α is open in $\mathbb{C}^m$ for each α. Then

(a) if $\boldsymbol{U}^T(\boldsymbol{z})\boldsymbol{I}(\boldsymbol{z})\boldsymbol{U}(\boldsymbol{z})$ is regular, there exists a strongly identifiable open neighborhood about $\boldsymbol{z}$, and

(b) if there exists a representative mapping $\boldsymbol{\varphi}^*_\alpha : \Omega_\alpha \to \mathbb{C}^q$ for each α, then the matrix $\boldsymbol{U}^T(\boldsymbol{z})\boldsymbol{I}(\boldsymbol{z})\boldsymbol{U}(\boldsymbol{z})$ is regular for every $\boldsymbol{z} \in \Omega$.

Proof. By the implicit function theorem (Theorem 3.3), then for any matrix $\boldsymbol{U}(\boldsymbol{\theta})$ whose columns form a basis for the null space of the Jacobian of $\boldsymbol{f}(\boldsymbol{\theta})$, there exists an open set $\mathbb{V} \ni \boldsymbol{\theta}$, an open set $\mathbb{W} \subset \mathbb{R}^{m-k}$, and some transformation $\boldsymbol{g_\theta} : \mathbb{W} \to \mathbb{R}^m$ such that $\boldsymbol{\theta} = \boldsymbol{g_\theta}(\boldsymbol{\xi})$ for some $\boldsymbol{\xi} \in \mathbb{W}$. In particular, in this reduced parameter space, there exists a FIM such that $\tilde{\boldsymbol{I}}(\boldsymbol{\xi}) = \boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$. Since Theorem 2.6 applies to $\tilde{\boldsymbol{I}}(\boldsymbol{\xi})$, the result for $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is proven. □

Regardless of the regularity of the FIM, only regularity of $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ determines strict identifiability under constraints for normal distributions, given a proper holomorphic function(s).

3.3 Linear Model

Assume the observations $\boldsymbol{x}$ model a linear function of the parameters $\boldsymbol{\theta}$, as in

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}, \tag{3.23}$$

where $\boldsymbol{H}$ is a full column rank $n \times m$ observation matrix consisting of known elements and $\boldsymbol{w}$ is a random noise vector with mean zero and known variance $\boldsymbol{C}$. As noted in section 2.3.1, the (weighted) LSE

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x}$$

is the BLUE. The LSE has a variance of $\boldsymbol{Q}^{-1}$ where $\boldsymbol{Q} = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)$. In addition, we assume a linear constraint

$$\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{F}\boldsymbol{\theta} + \boldsymbol{v} = \boldsymbol{0},$$

where $\boldsymbol{F}$ is a known full row rank $k \times m$ projection matrix and $\boldsymbol{v}$ is a known shift vector. For a linear constraint, the Jacobian $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{F}$ does not depend on the parameter. As the linear problem is well-studied, many of the results in this section are known (e.g., see [61, section 3.8] and [57, section 11.3.3]) but are presented here from a different perspective.

3.3.1 Best Linear Unbiased Estimation

The constrained (weighted) LSE (CLSE) is most often given by [36, p. 252]

$$\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x}) = \hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) - \boldsymbol{Q}^{-1}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{Q}^{-1}\boldsymbol{F}^T\right)^{-1}\left(\boldsymbol{F}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) + \boldsymbol{v}\right). \tag{3.24}$$

Simple calculation confirms that this CLSE exists in $\ker(\boldsymbol{f}) = \Theta_f$ (i.e., it satisfies the constraint), is unbiased, and has variance

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x})) \;=\; \boldsymbol{Q}^{-1} - \boldsymbol{Q}^{-1}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{Q}^{-1}\boldsymbol{F}^T\right)^{-1}\boldsymbol{F}\boldsymbol{Q}^{-1}. \tag{3.25}$$

The BLUE property of the LSE is preserved for the CLSE. This is not a surprising result since a linear constraint on a linear model is still a linear model, as seen in figure 3.2.

In the context of the null space approach of the CCRB, an alternative CLSE may be developed. One particular advantage of this approach is the avoidance of

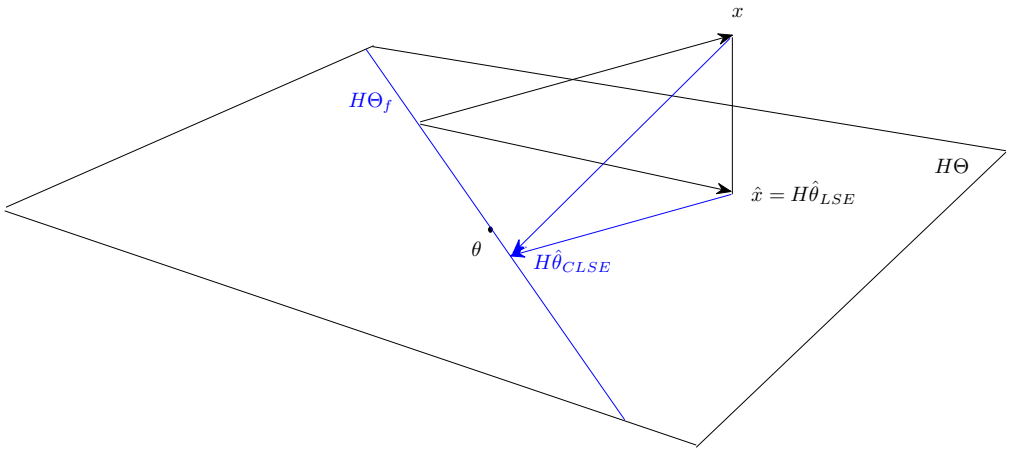


Figure 3.2: Projection of the observations $\boldsymbol{x}$ onto the linear space $\boldsymbol{H}\Theta$ and the linear constraint space $\boldsymbol{H}\Theta_f$.

the need for directly using and solving for Lagrange multipliers. Note that zero solutions of $\boldsymbol{f}(\boldsymbol{\theta})$ are of the form

$$\boldsymbol{\theta} = -\boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1} \boldsymbol{v} + \boldsymbol{U}\boldsymbol{\xi} = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$$

where $\boldsymbol{U}$ satisfies the equations in (3.6), i.e., the columns of $\boldsymbol{U}$ form an orthonormal basis for the null space of the row vectors of $\boldsymbol{F}$, and $\boldsymbol{\xi} \in \mathbb{R}^{m-k}$ is a parameter representing the projection of $\boldsymbol{\theta}$ to the constraint space Θ_f. Since the Jacobian $\boldsymbol{F}$ is independent of the parameter, then $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{U}$ and, hence, $\boldsymbol{g}_{\boldsymbol{\theta}}$ are as well. Moreover, the typical local properties for the implicit function hold globally. Substituting this solution for $\boldsymbol{\theta}$ the linear model is reformulated

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{w} \tag{3.26}$$

where $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{H}\boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1} \boldsymbol{v}$. Following the least squares result for the CRB,

we desire a solution that minimizes a quadratic objective function

$$\hat{\boldsymbol{\xi}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\xi}} \left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi}\right)^T \boldsymbol{C}^{-1} \left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi}\right).$$

This solution must satisfy the normal equations given by

$$\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)\hat{\boldsymbol{\xi}}_{\mathrm{LS}}(\boldsymbol{y}) = \boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{y}.$$

Provided the normalizing matrix $\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)$ is full rank, the LSE of $\boldsymbol{\xi}$ is given by

$$\hat{\boldsymbol{\xi}}_{\mathrm{LS}}(\boldsymbol{y}) \;=\; \left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{y}$$

and is the BLUE ((2.4) in section 2.3.1). The corresponding LSE of $\boldsymbol{\theta}$ based on this null space approach is

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x}) &= -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} + \boldsymbol{U}\hat{\boldsymbol{\xi}}_{\mathrm{LS}}(\boldsymbol{y}) \\
&= -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v}\right) \\
&= \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\left(\boldsymbol{\theta}_1 + \boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v}\right) - \boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} \\
&\qquad + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_1\right) \\
&= \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_1\right) \qquad (3.27)
\end{aligned}$$

where $\boldsymbol{\theta}_1 = -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} + \boldsymbol{U}\boldsymbol{\xi}_1$ can be any arbitrary point satisfying the linear constraint ($\boldsymbol{\xi}_1$ is unrestricted). This alternative CLSE is more general than the prior formula as it is applicable in scenarios when $\boldsymbol{H}$ and $\boldsymbol{F}$ are not necessarily full column rank and full row rank, respectively. As such, it has the more general expression for its variance

$$\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T,$$

which is equivalent to (3.25) when $\boldsymbol{H}$ and $\boldsymbol{F}$ are full column rank and full row rank, respectively. Replacing the rank conditions on $\boldsymbol{H}$ and $\boldsymbol{F}$, the weaker necessary conditions for this CLSE are that $\boldsymbol{HU}$ be full column rank. If the stronger necessary conditions on $\boldsymbol{H}$ and $\boldsymbol{F}$ exist, then the CLSE can be reformulated in terms of the LSE as before, e.g.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x}) = \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}) - \boldsymbol{\theta}_1\right).$$

It is easy to confirm that $\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x}) \in \ker(\boldsymbol{f})$ by recalling that $\boldsymbol{FU} = \boldsymbol{0}$ and $\boldsymbol{\theta}_1 \in \ker(\boldsymbol{f})$.

Also, as in section 2.3.1, if $\boldsymbol{HU}$ is not full column rank, then for estimable functions $\boldsymbol{d}^T\boldsymbol{\theta}$, i.e., for vectors $\boldsymbol{d}$ in the column space of $\boldsymbol{U}^T\boldsymbol{H}^T$, the LSE is BLUE and is given by

$$\boldsymbol{d}^T\hat{\boldsymbol{\theta}}_{CLS}(\boldsymbol{x}) = \boldsymbol{d}^T\boldsymbol{\theta}_1 + \boldsymbol{d}^T\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{\dagger}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_1\right) \tag{3.28}$$

similar to (3.27) with variance $\boldsymbol{d}^T\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{\dagger}\boldsymbol{U}^T\boldsymbol{d}$.

3.3.2 Uniform Minimum Variance Estimation under Gaussian noise

Under the assumption that the noise is normally distributed, the (unconstrained) LSE is also the MLE. The FIM[12] is $\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{Q} = \left(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\right)$, and the LSE/MLE is the MVUE being efficient with respect to CRB $= \boldsymbol{I}^{-1}(\boldsymbol{\theta})$. Given this general principle that for linear models with additive Gaussian noise, the LSE is the MLE, then since a linear constraint is essentially a reduced dimensional linear

[12]Note that neither the Fisher information, the Jacobian of the constraints, nor the null space matrix depend on the parameters in the linear model with linear constraints.

model as evidenced in (3.26), the CLSE should be the constrained MLE (CMLE). As we shall see, this is indeed the case.

First, for the Gaussian linear model in (3.26), the pdf is

$$q(\boldsymbol{y};\boldsymbol{\xi}) = \frac{1}{(2\pi)^{(m-k)/2}(\det \boldsymbol{C})^{(m-k)/2}} \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi}\right)^T \boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi}\right)\right\}.$$

The Fisher score $\tilde{\boldsymbol{s}}(\boldsymbol{y};\boldsymbol{\xi}) = -\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{U}\boldsymbol{\xi}\right)$ has a variance, or Fisher information, of $\tilde{\boldsymbol{I}}(\boldsymbol{\xi}) = \boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{U}$. Hence the CCRB is

$$\begin{aligned} \mathrm{CCRB}(\boldsymbol{\theta}) &= \boldsymbol{G}(\boldsymbol{\xi})\tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi})\boldsymbol{G}^T(\boldsymbol{\xi}) \\ &= \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T \end{aligned}$$

(see example 3.20).

Maximizing the likelihood (or pdf) is equivalent to minimizing the quadratic, therefore the LSE of $\boldsymbol{\xi}$ is also the MLE of $\boldsymbol{\xi}$. And by the invariance property [36, 14] of the MLE, then the CMLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) = \boldsymbol{g}(\hat{\boldsymbol{\xi}}_{\mathrm{ML}}(\boldsymbol{y})) = \hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x}).$$

Theorem 3.30. The CMLE is optimal for the linear model under linear constraints. That is, if the observations obey the linear model in (3.23), where $\boldsymbol{H}$ is a known matrix, $\boldsymbol{\theta}$ is an unknown parameter vector subject to the linear constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{F}\boldsymbol{\theta} + \boldsymbol{v} = \boldsymbol{0}$, and $\boldsymbol{w}$ is a zero-mean normal random vector with known variance $\boldsymbol{C}$, then provided $\boldsymbol{H}\boldsymbol{U}$ is full column rank, where $\boldsymbol{U}$ is defined by (3.6), the CMLE

$$\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) = \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}_1\right) \tag{3.29}$$

is unbiased and efficient.

While the formula for the CMLE based on the MLE might seem preferable, analogous to the formula for the CLSE based on the LSE, i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) = \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\left(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x}) - \boldsymbol{\theta}_1\right), \tag{3.30}$$

with $\boldsymbol{Q} = \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}$, this formulation requires the existence of a full rank FIM $\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{Q}$ in the MLE. The benefit of using the CMLE in (3.29) versus the CMLE in (3.30) is that the following proof does not require this assumption.

Proof. First note for any $\boldsymbol{\theta}, \boldsymbol{\theta}_1$ satisfying the constraints

$$\begin{aligned}
\boldsymbol{\theta} - \boldsymbol{\theta}_1 &= -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v} - \boldsymbol{U}\boldsymbol{\xi}_1 \\
&= \boldsymbol{U}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_1\right)
\end{aligned}$$

for some $\boldsymbol{\xi}, \boldsymbol{\xi}_1 \in \mathbb{R}^{m-k}$. Therefore, the expected value of the CMLE is

$$\begin{aligned}
E_{\boldsymbol{\theta}}\,\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) &= \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_1\right) \\
&= \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_1\right) \\
&= \boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_1\right) \\
&= \boldsymbol{\theta}_1 + \boldsymbol{\theta} - \boldsymbol{\theta}_1 \\
&= \boldsymbol{\theta}.
\end{aligned}$$

Finally, the variance of the CMLE is

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x})) &= \mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_1 + \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_1\right)\right) \\
&= \mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{H}\boldsymbol{\theta} - \boldsymbol{H}\boldsymbol{\theta}_1\right)\right) \\
&= \mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\right)\right) \\
&= \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T \\
&= \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T \\
&= \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T,
\end{aligned}$$

i.e., the CCRB of $\boldsymbol{\theta}$. □

Thus, the CMLE is the MVUE for the linear model with linear constraints under a Gaussian assumption.

Additionally, when $\boldsymbol{H}\boldsymbol{U}$ is not full column rank, then when $\boldsymbol{d}$ is in the column space of $\boldsymbol{U}^T\boldsymbol{H}^T$, the MLE of $\boldsymbol{d}^T\boldsymbol{\theta}$ is still the MVUE and is given by $\boldsymbol{d}^T\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) = \boldsymbol{d}^T\hat{\boldsymbol{\theta}}_{\mathrm{CLS}}(\boldsymbol{x})$ from (3.28), with variance $\boldsymbol{d}^T\mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{d}$.

3.4 Constrained Maximum Likelihood Estimation

The constrained MLE (CMLE) of the parameter vector $\boldsymbol{\theta}$ constrained to the manifold $\Theta_f = \{\boldsymbol{\theta}' : \boldsymbol{f}(\boldsymbol{\theta}') = \boldsymbol{0}\}$ is the estimator in Θ_f, which given the observations $\boldsymbol{x}$, that maximizes the likelihood distribution $p(\boldsymbol{x};\cdot)$, i.e. it is the maximum likelihood in Θ_f. Since $\log(\cdot)$ is concave, it is convenient to equivalently maximize the log-likelihood $\log p(\boldsymbol{x};\cdot)$ since then the Jacobian of the objective is the Fisher score. In an optimization context, the CMLE, which will be denoted $\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x})$, is

the solution to the following constrained optimization problem:

$$\begin{array}{ll} \max\limits_{\boldsymbol{\theta}'} & \log p(\boldsymbol{x};\boldsymbol{\theta}') \\ \text{s.t.} & \boldsymbol{f}(\boldsymbol{\theta}') = \mathbf{0}. \end{array} \tag{3.31}$$

Analogous to the method of maximum likelihood approach of (2.7), solutions $\dot{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{g}_{\dot{\boldsymbol{\theta}}(\boldsymbol{x})}(\dot{\boldsymbol{\xi}}(\boldsymbol{x}))$ satisfying

$$\frac{\partial}{\partial \boldsymbol{\xi}'} \log p(\boldsymbol{x}, \boldsymbol{g}_{\boldsymbol{\theta}'}(\boldsymbol{\xi}')) \Big|_{\boldsymbol{\xi}'=\dot{\boldsymbol{\xi}}(\boldsymbol{x})} = \mathbf{0},$$

where $\boldsymbol{g}_{\boldsymbol{\theta}'}$ is defined by (3.7), are candidates to be the CMLE. More formally, a solution to this optimization problem must satisfy the Karush-Kuhn-Tucker conditions [45], i.e.,

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}') - \boldsymbol{\lambda}^T \boldsymbol{F}(\boldsymbol{\theta}') = \mathbf{0} \quad \text{(stationarity)} \tag{3.32}$$

$$\boldsymbol{f}(\boldsymbol{\theta}') = \mathbf{0}. \quad \text{(feasibility)} \tag{3.33}$$

Any point satisfying these conditions is a *stationary, feasible point.*

Since $\hat{\boldsymbol{\theta}}_{\text{CML}}(\boldsymbol{x}) \in \Theta_f$, then the implicit function theorem implies there exists an open set $\mathbb{O} \subset \Theta_f$ containing $\hat{\boldsymbol{\theta}}_{\text{CML}}(\boldsymbol{x})$, an open set $\mathbb{P} \subset \mathbb{R}^{m-k}$, and a continuously differentiable bijection $\boldsymbol{g}_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})} : \mathbb{P} \to \mathbb{O}$ such that $\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{g}_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}(\hat{\boldsymbol{\xi}}(\boldsymbol{x}))$ for some $\hat{\boldsymbol{\xi}}(\boldsymbol{x}) \in \mathbb{P}$. If $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is a maximizer of the likelihood function $p(\boldsymbol{x};\boldsymbol{\theta})$ in the constraint set Θ_f, then the likelihood $q(\boldsymbol{x};\boldsymbol{\xi}') = p(\boldsymbol{x};\boldsymbol{g}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\xi}'))$ has a maximum at $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$ in $\mathbb{P}$ (i.e., a local maximum at $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$ in $\mathbb{R}^{m-k}$). It cannot be said that $q(\boldsymbol{x};\boldsymbol{\xi})$ has a global maximum at $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$, since $\boldsymbol{\theta}' = \boldsymbol{g}_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}(\boldsymbol{\xi}')$ is only guaranteed to exist in Θ_f when $\boldsymbol{\xi}' \in \mathbb{P}$, i.e., there may exist a point $\boldsymbol{\xi}' \in \mathbb{R}^{m-k} \backslash \mathbb{P}$ such that $q(\boldsymbol{x};\boldsymbol{\xi}') > q(\boldsymbol{x};\hat{\boldsymbol{\xi}}(\boldsymbol{x}))$ and $\boldsymbol{g}_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}(\boldsymbol{\xi}') \notin \Theta_f$.

3.4.1 Efficient estimation

In Section 2.4.1, it was explained that when an efficient estimator exists, the method of maximum likelihood finds the estimator [36, 62]. It is useful to note the connection between efficiency and the method of constrained maximum likelihood since Stoica and Ng ignored this extension in their paper [68], despite Marzetta having showed that this result extends to the constrained case when the FIM is non-singular [47, theorem 3]. What follows is the general extension of this result, including the case for singular FIMs.

Theorem 3.31. If $\boldsymbol{t}(\boldsymbol{x})$ is a constrained estimator of $\boldsymbol{\theta}$, required to satisfy the constraint $\boldsymbol{f}(\boldsymbol{t}(\boldsymbol{x})) = \mathbf{0}$, which is also efficient with respect to the CCRB, then the estimator is a stationary point for the constrained optimization problem in (3.31).

Proof. This is perhaps more easily proven strictly from the constrained parameter perspective, since the global maximum of the likelihood relative to the implicit reparameterization may not correspond to global maximum in Θ_f relative to the constrained parameterization. Since $\boldsymbol{t}(\boldsymbol{x})$ is efficient then in the mean-square sense we have $\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} = \mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. Then as $\boldsymbol{\theta} \to \boldsymbol{t}(\boldsymbol{x})$ (this assumes the observations are consistent with $\boldsymbol{\theta}$) we have

$$\mathbf{0} \leftarrow \boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\theta} = \mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}).$$

The continuity of the CCRB and the Fisher score implies $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{t}(\boldsymbol{x})) = \boldsymbol{F}^T(\boldsymbol{t}(\boldsymbol{x})) \cdot \boldsymbol{\lambda}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^k$ or

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{t}(\boldsymbol{x})) - \boldsymbol{\lambda}^T\boldsymbol{F}(\boldsymbol{t}(\boldsymbol{x})) = \mathbf{0},$$

which defines the stationarity condition (3.32) of the constrained optimization problem with $\boldsymbol{\lambda}$ being the vector of Lagrange multipliers. □

3.4.2 Asymptotic Normality

The asymptotic properties of the MLE can be found in section 2.4.2. Therein, it was mentioned that the maximum likelihood estimator was asymptotically unbiased and efficient with variance asymptotically equivalent to the CRB. A corresponding relationship exists between the CMLE and the CCRB. As before, let the samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, be iid as $\boldsymbol{x}$ from the likelihood $p(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is assumed to exist in Θ_f. Denote $\boldsymbol{y}_n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ to be the collection of these samples, so that the likelihood will be $p(\boldsymbol{y}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta})$. Hence, the asymptotic CMLE will be denoted $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$.

Theorem 3.32. Assuming the pdf satisfies the regularity conditions (see (3.19) and discussion after proof), then the CMLE is asymptotically distributed according to

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathrm{CCRB}(\boldsymbol{\theta})\right).$$

There exists a number of results in the literature regarding the asymptotic characteristics of the CMLE (e.g., the works of Aitchison and Silvey [3, 63, 4, 2], and of Crowder [18]). For example, Crowder shows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) - \boldsymbol{\theta}\right) \xrightarrow{d}$$
$$\mathcal{N}\left(\mathbf{0}, \boldsymbol{D}^{-1}(\boldsymbol{\theta}) - \boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\right)$$

where $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{K}\boldsymbol{F}(\boldsymbol{\theta})$ for any positive semi-definite matrix $\boldsymbol{K}$ that ensures the nonsingularity of $\boldsymbol{D}(\boldsymbol{\theta})$. And while it would be sufficient to use these

existing results to verify the connection between the CMLE and the CCRB, it is also insightful (and the point of this treatise) to examine the problem entirely from the perspective of the reduced parameter space, i.e., using the implicit function or a null space approach.[13]

Proof. By the implicit function theorem (Theorem 3.3), there exists an open set $\mathbb{O} \subset \Theta_f$ containing $\boldsymbol{\theta}$, an open set $\mathbb{P} \subset \mathbb{R}^{m-k}$, and a continuously differentiable bijection $\boldsymbol{g}_{\boldsymbol{\theta}} : \mathbb{P} \to \mathbb{O}$ such that $\boldsymbol{\theta} = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$ for some $\boldsymbol{\xi} \in \mathbb{P}$. The likelihood for $\boldsymbol{\xi}$ is given by $q(\boldsymbol{y}_n; \boldsymbol{\xi}) = p(\boldsymbol{y}_n; \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}))$.

Let $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)$ be the MLE of $\boldsymbol{\xi}$ based on the likelihood $q(\boldsymbol{y}_n; \boldsymbol{\xi})$. Since the MLE is consistent and asymptotically efficient, then

$$\sqrt{n}\left(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) - \boldsymbol{\xi}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi})\right). \tag{3.34}$$

In particular, since $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) \to \boldsymbol{\xi}$ as $n \to \infty$, then for n sufficiently large, say $n > N$, $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) \in \mathbb{P}$. Let $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$ be the CMLE of $\boldsymbol{\theta}$ based on the likelihood $p(\boldsymbol{y}_n; \boldsymbol{\theta})$ and the constraint $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$. By the invariance property [36, 62], for $n > N$, $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) = \boldsymbol{g}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) \in \mathbb{O}$. Therefore, since $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) \to \boldsymbol{\xi}$ as $n \to \infty$, $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) \to \boldsymbol{\theta}$ also and the CMLE is consistent.

The Taylor series expansion (see section 3.1.5) of $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}^{'})$ can be truncated using a Lagrange remainder term [38, p. 232] as in

$$\boldsymbol{g}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)) = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}) + \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))\left(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) - \boldsymbol{\xi}\right)$$

where $\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))$ is meant to represent a matrix of the form of $\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$ where each

[13] Nevertheless, a proof directly from Crowder's asymptotic normality result is detailed in appendix A.3.

row is evaluated at possibly different points $\boldsymbol{\xi}'^{(i)}$, $i = 1, \ldots, m-k$, each existing on the line segment starting at $\boldsymbol{\xi}$ and ending with $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)$. From the invariance property of the MLE, this can be rewritten as

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) - \boldsymbol{\theta} = \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)) \left(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n) - \boldsymbol{\xi} \right).$$

Since the MLE $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)$ is consistent, then $\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)) \xrightarrow{p} \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. Given this and (3.34), then by Slutsky's theorem [62, p. 60]

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n) - \boldsymbol{\theta} \right) \xrightarrow{d} \mathcal{N} \left(\boldsymbol{0}, \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\xi}) \tilde{\boldsymbol{I}}^{-1}(\boldsymbol{\xi}) \boldsymbol{G}_{\boldsymbol{\theta}}^T(\boldsymbol{\xi}) \right),$$

which by theorem 3.5 shows $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$ is asymptotically efficient with respect to the CCRB. □

The conditions for asymptotic normality with respect to the CCRB are the conditions that $\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)$ be asymptotically normal [14, p. 516]. For the MLE these include (a) differentiability of the Fisher score, (b) the Fisher information continuous with respect to the parameter and nonzero at $\boldsymbol{\xi}$, and (c) consistency. For the CMLE and theorem 3.32, this translates to (a) differentiability of the Fisher score and the existence of first and second derivatives of any implicit function (or equivalently, the constraint $\boldsymbol{f}$), (b) $\boldsymbol{U}^T(\boldsymbol{\theta}')\boldsymbol{I}(\boldsymbol{\theta}')\boldsymbol{U}(\boldsymbol{\theta}')$ continuous with respect to $\boldsymbol{\theta}'$ and regular at $\boldsymbol{\theta}' = \boldsymbol{\theta}$, and (c) consistency of the CMLE.

3.4.3 The Method of Scoring Under Parametric Constraints

The method of scoring for unconstrained parameters is detailed in section 2.4.3. Here, we examine scoring with constraints. Assume we have an iterate $\dot{\boldsymbol{\theta}}^{(k)} \in \Theta_f$ and

we wish to improve this iterate in the sense of the optimization problem expressed in (3.31). The method of scoring does not directly apply, since any projection step will not take into account the constraint, i.e., it is likely the direction of steepest ascent is not the appropriate path in terms of maximizing the likelihood subject to the functional equality constraints. Thus, it is desirable to have projected direction and restoration steps that take the constraints into consideration.

Given an initial estimate $\dot{\boldsymbol{\theta}}^{(k)}$, there exists a set $\mathbb{O} \ni \dot{\boldsymbol{\theta}}^{(1)}$ open in Θ_f, a set $\mathbb{P}$ open in $\mathbb{R}^{m-k}$, and a continuously differentiable function $\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}} : \mathbb{R}^{m-k} \rightarrow \mathbb{R}^m$ such that $\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}$ is a diffeomorphism on $\mathbb{P}$, $\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\mathbb{P}) = \mathbb{O}$, and in particular there exists a $\dot{\boldsymbol{\xi}}^{(k)} \in \mathbb{P}$ such that $\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) = \dot{\boldsymbol{\theta}}^{(k)}$. Scoring can now be applied in the reduced parameter space of $\mathbb{R}^{m-k}$.

For the given set of observations $\boldsymbol{x}$ and this corresponding initial estimate $\dot{\boldsymbol{\xi}}^{(k)}$, the method of scoring suggests the projection step

$$\dot{\boldsymbol{\xi}}^{(k+1)} = \dot{\boldsymbol{\xi}}^{(k)} + \tilde{\boldsymbol{I}}^{-1}(\dot{\boldsymbol{\xi}}^{(k)})\tilde{\boldsymbol{s}}(\boldsymbol{x}; \dot{\boldsymbol{\xi}}^{(k)})$$

to generate a better estimate $\dot{\boldsymbol{\xi}}^{(k+1)}$ in the sense of maximizing the likelihood $q(\boldsymbol{x}; \boldsymbol{\xi}^{'})$ $= p(\boldsymbol{x}; \boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\boldsymbol{\xi}^{'}))$. As with many iterative procedures, convergence is only guaranteed under certain initial conditions. If the projection step or shift is too large, then $\dot{\boldsymbol{\theta}}^{(k+1)}$ may not be a *usable* point, i.e., an iterate that increases the value of the likelihood function. To add stability to the procedure, often a step-size rule or shift-cutting is employed. This amounts to the inclusion of a multiplicative factor $\alpha^{(k)} \in [0, 1]$, modifying the projection step to

$$\dot{\boldsymbol{\xi}}^{(k+1)} = \dot{\boldsymbol{\xi}}^{(k)} + \alpha^{(k)}\tilde{\boldsymbol{I}}^{-1}(\dot{\boldsymbol{\xi}}^{(k)})\tilde{\boldsymbol{s}}(\boldsymbol{x}; \dot{\boldsymbol{\xi}}^{(k)}).$$

Choosing an appropriate *step-size rule* for $\alpha^{(k)}$ will guarantee convergence, although typically at a cost to the rate of convergence.

The Taylor series expansion (see section 3.1.5) of $\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}$ about $\dot{\boldsymbol{\xi}}^{(k)}$ and evaluated at $\dot{\boldsymbol{\xi}}^{(k+1)}$ is given by

$$\boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k+1)}) = \boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) + \boldsymbol{G}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \cdot (\dot{\boldsymbol{\xi}}^{(k+1)} - \dot{\boldsymbol{\xi}}^{(k)}) + o(||\dot{\boldsymbol{\xi}}^{(k+1)} - \dot{\boldsymbol{\xi}}^{(k)}||)$$

where $o(||\dot{\boldsymbol{\xi}}^{(k+1)} - \dot{\boldsymbol{\xi}}^{(k)}||)$ is a term that shrinks faster than $||\dot{\boldsymbol{\xi}}^{(k+1)} - \dot{\boldsymbol{\xi}}^{(k)}||$ as $k \to \infty$. Ignoring this error term, this generates an iteration in the larger dimensional parameter space $\Theta \subset \mathbb{R}^m$ by defining the next iterate $\dot{\boldsymbol{\theta}}^{(k+1)} = \boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k+1)})$. That is,

$$\begin{aligned}
\dot{\boldsymbol{\theta}}^{(k+1)} &= \boldsymbol{g}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) + \boldsymbol{G}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \cdot (\dot{\boldsymbol{\xi}}^{(k+1)} - \dot{\boldsymbol{\xi}}^{(k)}) \\
&= \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \boldsymbol{G}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \tilde{\boldsymbol{I}}^{-1}(\dot{\boldsymbol{\xi}}^{(k)}) \tilde{\boldsymbol{s}}(\boldsymbol{x}; \dot{\boldsymbol{\xi}}^{(k)}) \\
&= \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \boldsymbol{G}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \left(\boldsymbol{G}^T_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{G}_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \right)^{-1} \boldsymbol{G}^T_{\dot{\boldsymbol{\theta}}^{(k)}}(\dot{\boldsymbol{\xi}}^{(k)}) \boldsymbol{s}(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)}) \\
&= \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(k)}) \left(\boldsymbol{U}^T(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(k)}) \right)^{-1} \boldsymbol{U}^T(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{s}(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)}).
\end{aligned}$$

In comparison with the classical method of scoring, this iteration

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{s}(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)}) \tag{3.35}$$

is essentially a replacement of the CRB with the CCRB. This is the projection step of the method of scoring with parametric equality constraints.[14] Intuitively, this should seem appropriate since the CCRB is a generalization of the CRB. However,

[14]Osborne [55] used a Lagrangian multiplier approach to develop the method of scoring. But his scenario was restricted to linear constraints and, hence, lacked the restoration step. Additionally, he makes no mention that the matrix projecting the negative Jacobian of the objective is a constrained Cramér-Rao bound. Note the structure of this projection step is well-known as a nonstatistical formulation exists for the conventional optimization problem in [25, p. 178].

even with an appropriate step-size rule to generate usable iterates, since there is no certainty that $\dot{\boldsymbol{\xi}}^{(k+1)} \in \mathbb{P}$, then it is likely that $\dot{\boldsymbol{\theta}}^{(k+1)}$ will not be a feasible point. To correct this, an encompassing restoration step is required to produce the next iterate, i.e.,

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \boldsymbol{\pi}\left[\dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)}\mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k)})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\right] \tag{3.36}$$

where $\boldsymbol{\pi}[\cdot]$ is the natural projection of $\mathbb{R}^m$ onto Θ_f. This is the method of scoring with parametric equality constraints. With this additional restoration step, the

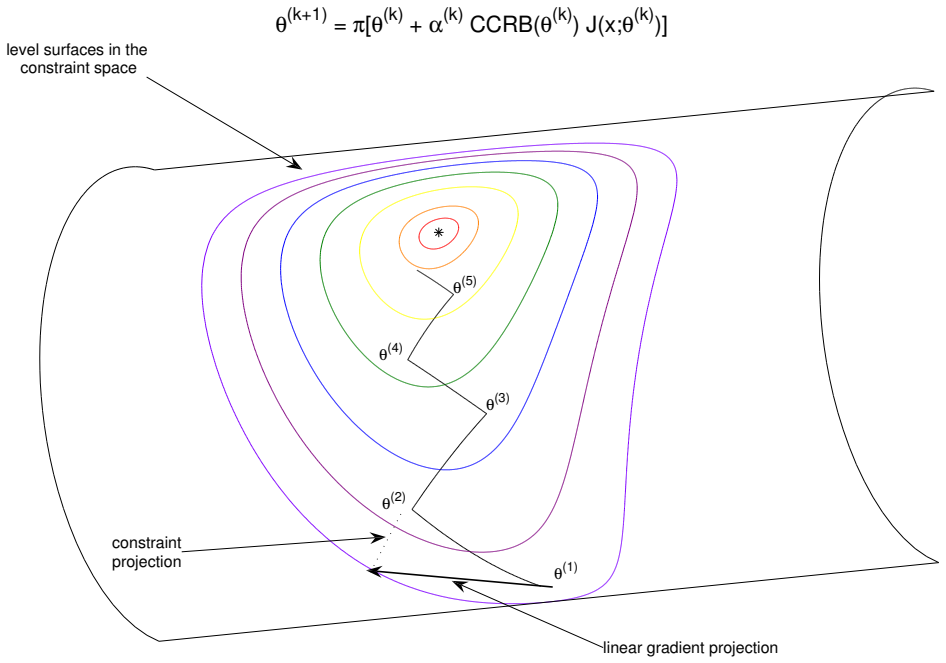


Figure 3.3: Path created by iterates from the method of scoring with constraints.

usability of the iterate would be tested and accepted or rejected after (3.36) as opposed to after (3.35). For a convex set, the natural projection is well-defined. In general, though, some other rule will likely need to be applied in cases for which there does not exist a unique shortest distance to Θ_f, e.g., reducing the step size

$\alpha^{(k)}$. Simple projections, e.g., onto planar or spherical constraints, are relatively simple operations, however, it might be more commonly the case that the restoration cannot be expressed analytically. To ensure the iterates satisfy the constraints approximately, one approach is to apply an additional iterative process [25]

$$\dot{\boldsymbol{\theta}}^{(k,l+1)} = \dot{\boldsymbol{\theta}}^{(k,l)} - \boldsymbol{F}^T(\dot{\boldsymbol{\theta}}^{(k,l)}) \left(\boldsymbol{F}(\dot{\boldsymbol{\theta}}^{(k,l)})\boldsymbol{F}^T(\dot{\boldsymbol{\theta}}^{(k,l)})\right)^{-1} \boldsymbol{f}(\dot{\boldsymbol{\theta}}^{(k,l)}),$$

where $\dot{\boldsymbol{\theta}}^{(k,1)} = \dot{\boldsymbol{\theta}}^{(k)}$ and $\dot{\boldsymbol{\theta}}^{(k+1)} = \boldsymbol{\pi}\left[\dot{\boldsymbol{\theta}}^{(k)}\right] = \dot{\boldsymbol{\theta}}^{(k,l)}$ when $\boldsymbol{f}(\dot{\boldsymbol{\theta}}^{(k,l)}) \approx \mathbf{0}$ to a desired degree of accuracy and provided the iterate is still usable. Alternatively, a penalty can be added to the cost (objective) function, e.g., as in

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}') + \eta \sum_{i=1}^{k} |f_i(\boldsymbol{\theta}')|$$

for some positive η and where f_i is the ith constraint equation, to limit the divergence of the iterations away from Θ_f.

3.4.3.1 Convergence Properties

There is a large class of conditions that guarantee convergence in fixed point theorems, some of which can be found in [64, 25, 10]. The most general statement is that given an initialization "sufficiently close" to the maximum value $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$, the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$ generated by the algorithm will converge to this CMLE. As the method of scoring with parametric equality constraints is a Newton-type method, convergence properties that already exist for these methods can be adapted here. As it is impossible to cover all the potential approaches to developing properties for this constrained scoring algorithm, this section focuses on properties similar to those

found in Goldstein [22]. To ease reading of this section, the proofs of the theorems in this section are presented in appendix B.

First, define $\Theta_{\dot{\boldsymbol{\theta}}^{(k)}} = \{\boldsymbol{\theta}' \in \Theta_f : p(\boldsymbol{x}; \boldsymbol{\theta}') \geq p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$ as the set of all feasible and usable iterates after the kth iterate $\dot{\boldsymbol{\theta}}^{(k)}$. The step rule for the properties in this section is as follows: for a fixed $\beta \in (0, 1)$ choose the least positive integer $m^{(k)}$ such that $\alpha^{(k)} = \beta^{m^{(k)}}$ satisfies the inequality

$$\alpha^{(k)} \left(\log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k+1)}) - \log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)}) \right) \geq \kappa \left\| \dot{\boldsymbol{\theta}}^{(k+1)} - \dot{\boldsymbol{\theta}}^{(k)} \right\|^2_{\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})} \tag{3.37}$$

where $\dot{\boldsymbol{\theta}}^{(k+1)}$ is defined by (3.36). If no finite $m^{(k)}$ exists, then choose $\alpha^{(k)} = 0$. This type of step-size rule enforces a stepwise Lipschitz condition. For theorems 3.33 and 3.37, we require that Θ_f be convex.[15]

Theorem 3.33. If for any iterate $\dot{\boldsymbol{\theta}}^{(k)} \in \Theta_f$ there does not exist an $\alpha^{(k)} > 0$ that satisfies (3.37), then $\dot{\boldsymbol{\theta}}^{(k)}$ is a stationary point.

Therefore, when the step rule forces the choice of $\alpha^{(k)} = 0$ then the method of scoring with parametric equality constraints has converged. The next theorem details a property on the sequence of likelihood functions generated by the iterates.

Theorem 3.34. The sequence $\{p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$ is a monotone increasing sequence. Furthermore, if $p(\boldsymbol{x}; \cdot)$ is bounded above, then $\{p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$ converges.

[15] For any nonlinear equality constraint, Θ_f will not be convex. However, locally, the restoration of the linear projection onto the tangent space of the CCRB has the appearance of the restoration onto a convex set for sufficiently small $\alpha^{(k)}$, i.e., Θ_f appears locally convex. While it may be possible to restrict the step size to this local convexity, such an enhancement is beyond the scope of the work presented here. For inequality constraints, Θ_f may be convex; although such constraints, as in example 3.15, will not inform the projection update in (3.35), they might inform the restoration update in (3.36).

Thus, $\Theta_{\dot{\boldsymbol{\theta}}^{(k)}}$ is a decreasing sequence of closed (nested) sets, i.e., $\Theta_{\dot{\boldsymbol{\theta}}^{(k+1)}} \subset \Theta_{\dot{\boldsymbol{\theta}}^{(k)}}$ or for any given sequence, $\dot{\boldsymbol{\theta}}^{(i)} \in \Theta_{\dot{\boldsymbol{\theta}}^{(j)}}$ provided $i \geq j$. That is, using a proper step size rule will guarantee usable iterates. The monotonicity of $\{p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$, even if bounded above, does not imply monotonicity in the sequence $\{\log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k+1)}) - \log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$. However, this does guarantee convergence.

Theorem 3.35. If the likelihood $p(\boldsymbol{x}; \cdot)$ is bounded above, then the sequence

$$\{\log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k+1)}) - \log p(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})\}$$

vanishes.

Hence, a bounded likelihood function guarantees the existence of a maximum likelihood solution(s). This can also be shown by the nested interval theorem [38, problem 2-1-12]. These previous properties would be a consequence of any rule that chooses feasible, usable iterates. The value of the rule in (3.37) is that it allows for statements to be made on the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$.

Theorem 3.36. If the likelihood $p(\boldsymbol{x}; \cdot)$ is bounded above, then the sequence

$$\{\|\dot{\boldsymbol{\theta}}^{(k+1)} - \dot{\boldsymbol{\theta}}^{(k)}\|_{\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})}\}$$

vanishes as $k \to \infty$.

This theorem does not guarantee that the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$ will converge or even have a *limit point.*[16] (The series $\sum_{j=1}^{k} \frac{1}{j}$ is an example of a sequence satisfying the above theorem with no real-valued limit points.) Although, if the set $\Theta_{\dot{\boldsymbol{\theta}}^{(k)}}$ is

[16] A point a is a *limit point* of a sequence $\{a_n\}$ if for any integer K and any $\epsilon > 0$, there exists an $k > K$ such that $|a_k - a| < \epsilon$.

bounded for some k, then the Bolzano-Weierstrass theorem [38, p. 52, theorem 2-12] implies the existence of a limit point of the sequence.

Theorem 3.37. If $\Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$ is compact and convex, then limit points of the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$ are also stationary points.

The theorem remains true if $\Theta_{\dot{\boldsymbol{\theta}}^{(k)}}$ is compact and convex for any k.

Theorem 3.38. If $\Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$ is compact for all sequences in a closed set of Θ_f and if there is a unique limit point $\boldsymbol{\theta}^\star$ for all such sequences then $\lim_{k\to\infty} \dot{\boldsymbol{\theta}}^{(k)} = \boldsymbol{\theta}^\star$ for every sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$. Also, $\boldsymbol{\theta}^\star$ is the maximum of $p(\boldsymbol{x};\cdot)$.

3.4.3.2 Linear constraints

Linear constraints on the parameter typically restrict the parameter to a set of the form $\Theta_f = \{\boldsymbol{\theta}' \in \Theta : \boldsymbol{F}\boldsymbol{\theta}' + \boldsymbol{v} = \boldsymbol{0}\}$ (see section 3.3). Under this linear constraint, the restoration operation $\boldsymbol{\pi}[\cdot]$ is redundant since any step remains in the constraint space, i.e., since $\dot{\boldsymbol{\theta}}^{(k)} \in \Theta_f$ and $\boldsymbol{F} \cdot \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k)}) = \boldsymbol{0}$. Thus the method of scoring with parametric equality constraints in (3.36) simplifies to the iteration

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{s}(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)}).$$

Example 3.39 (Linear model with linear constraints)**.** In the linear model with normal noise case of section 3.3.2, we have $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$ with $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$. In this case, the negative Hessian is the FIM and the optimization problem becomes a null space quadratic exercise [21], i.e., the minimization of a quadratic objective subject to a linear constraint. The $\mathrm{CCRB} = \boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T$, which is constant with

respect to the parameter; and the Fisher score is $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}') = \boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}'\right)$. Hence, if $\dot{\boldsymbol{\theta}}^{(1)}$ is any feasible vector, e.g., $\dot{\boldsymbol{\theta}}^{(1)} = -\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1}\boldsymbol{v}$, then the method of scoring with constraints finds the CMLE in one step to be

$$\dot{\boldsymbol{\theta}}^{(2)} = \dot{\boldsymbol{\theta}}^{(1)} + \mathrm{CCRB}\cdot\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(1)}\right)$$

which is exactly the formula in (3.29) from theorem 3.30. The next iterate $\dot{\boldsymbol{\theta}}^{(3)} = \dot{\boldsymbol{\theta}}^{(2)} + \mathrm{CCRB}\cdot\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(2)}\right)$ reveals that the procedure reaches a fixed point since

$$\begin{aligned}
&\mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(2)}\right)\\
&= \mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\left(\dot{\boldsymbol{\theta}}^{(1)} + \mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(1)}\right)\right)\right)\\
&= \mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(1)}\right) - \mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\mathrm{CCRB}\boldsymbol{H}^T\boldsymbol{C}^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\dot{\boldsymbol{\theta}}^{(1)}\right)\\
&= \boldsymbol{0}
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\theta}}_{\mathrm{CML}}(\boldsymbol{x}) = \dot{\boldsymbol{\theta}}^{(2)}$.

Example 3.40 (Jamshidian's GP algorithm)**.** Jamshidian [33] developed a Gradient Projection (GP) algorithm for maximizing the likelihood subject to linear parameter constraints using the iteration

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)}\left(\boldsymbol{W}^{-1} - \boldsymbol{W}^{-1}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{W}^{-1}\boldsymbol{F}^T\right)^{-1}\boldsymbol{F}\boldsymbol{W}^{-1}\right)\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)}) \qquad (3.38)$$

for some positive definite matrix $\boldsymbol{W}$. An optimal choice with regard to the algorithm's rate of convergence, Jamshidian suggests, is a possibly diagonally loaded Hessian of the log-likelihood

$$\boldsymbol{W}(\boldsymbol{x},\dot{\boldsymbol{\theta}}^{(k)}) = \frac{\partial}{\partial\boldsymbol{\theta}^T}\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)}) + \gamma^{(k)}\boldsymbol{I}_{m\times m},$$

where $\gamma^{(k)} \geq 0$ is chosen to be sufficiently large enough to ensure the positive definiteness of the matrix $\boldsymbol{W}$. This formulation is closely connected to the method of scoring with constraints. The GP iteration is equivalent to scoring by choosing $\boldsymbol{W}(\boldsymbol{x}, \dot{\boldsymbol{\theta}}^{(k)})$ to be the FIM, when it is nonsingular. Indeed, the projecting matrix is similar to the Marzetta form of the CCRB in [47] with $\boldsymbol{W}(\boldsymbol{x}, \dot{\boldsymbol{\theta}}^{(k)})$ replacing the FIM. This fact produces a slight generalization of the GP iteration, given by

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \boldsymbol{U} \left(\boldsymbol{U}^T \boldsymbol{W}(\boldsymbol{x}, \dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{U} \right)^{-1} \boldsymbol{U}^T \boldsymbol{s}(\boldsymbol{x}; \dot{\boldsymbol{\theta}}^{(k)})$$

where $\boldsymbol{U}$ is defined as in (3.6). In this formulation, the projecting metric $\boldsymbol{W}(\boldsymbol{x}, \dot{\boldsymbol{\theta}}^{(k)})$ only needs to be positive semidefinite. Alternatively, in (3.38) the Aitchison and Silvey [4] substitution for the FIM, $\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)}) + \boldsymbol{F}^T(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{K} \boldsymbol{F}(\dot{\boldsymbol{\theta}}^{(k)})$, instead of a diagonally loaded Hessian of the log-likelihood (or even a diagonally loaded Fisher information matrix).

In this sense, the two iterations are equivalent for the linear model, when the FIM is simply the negative Hessian. This occurs when the log-likelihood is quadratic (normal). This also suggests the adaption of Jamshidian's GP algorithm to cases of nonlinear constraints. Likewise, asymptotically, where $\boldsymbol{y}_n$ is denoted as in section 3.4.2, then by the law of large numbers the Jamshidian projection matrix

$$\boldsymbol{W}(\boldsymbol{y}_n, \dot{\boldsymbol{\theta}}^{(k)}) \rightarrow n \boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)}) + \boldsymbol{\gamma}^{(k)} \boldsymbol{I}_{m \times m}$$

for an arbitrarily small (possibly zero) $\boldsymbol{\gamma}^{(k)} > 0$, as $\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})$ is positive semidefinite. The projection step update then becomes

$$\dot{\boldsymbol{\theta}}^{(k+1)} = \dot{\boldsymbol{\theta}}^{(k)} + \alpha^{(k)} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k)}) \boldsymbol{s}(\boldsymbol{y}_n; \dot{\boldsymbol{\theta}}^{(k)}) - \alpha^{(k)} \gamma^{(k)} \boldsymbol{s}(\boldsymbol{y}_n; \dot{\boldsymbol{\theta}}^{(k)}),$$

i.e., essentially equivalent to the method of scoring with parametric equality constraints in (3.35).

3.5 Hypothesis testing

In section 2.5, hypothesis testing (the Rao and Wald tests) using the CRB was reviewed. In this section, hypothesis testing is considered under a constrained alternative. Assume $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^r$ is a consistent and nonredundant differentiable function, which is also consistent and nonredundant with the differentiable function $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^k$. Hence, $\Theta_h = \{\boldsymbol{\theta}' : \boldsymbol{h}(\boldsymbol{\theta}') = \mathbf{0}\} \subset \Theta_f$, where Θ_f is defined as in (3.4), and also $\text{rank}(\begin{bmatrix}\boldsymbol{H}(\boldsymbol{\theta}) \\ \boldsymbol{F}(\boldsymbol{\theta})\end{bmatrix}) = r + k \leq m$ and $r < m - k$. Then the hypothesis test can be stated as

$$H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}. \tag{3.39}$$

Naturally, $\boldsymbol{f}(\boldsymbol{\theta}') = \mathbf{0}$ under these conditions defines an implicit function locally, so assume $\boldsymbol{g}_{\boldsymbol{\theta}'} : \mathbb{R}^{m-k} \to \mathbb{R}^k$ is such a function satisfying theorem 3.3 for any $\boldsymbol{\theta}' \in \Theta_f$. Then, a locally (or asymptotically) equivalent hypothesis can be stated as

$$H_0 : \boldsymbol{h}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})) = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{h}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi})) \neq \mathbf{0}. \tag{3.40}$$

In this formulation, the well-known Rao and Wald statistics were shown in section 2.5.

3.5.1 The Rao statistic

For the hypothesis testing scenario in (3.40), the Rao test statistic presented in section 2.5.1 is given by

$$\rho(\boldsymbol{y}_n) = \tilde{\boldsymbol{s}}^T(\boldsymbol{y}_n; \hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))\tilde{\boldsymbol{I}}_n^{-1}(\hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))\tilde{\boldsymbol{s}}(\boldsymbol{y}_n; \hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))$$

where $\tilde{\boldsymbol{s}}(\boldsymbol{y}_n; \hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))$ is the Fisher score of the observations $\boldsymbol{y}_n$ (as defined in section 3.4.2) and evaluated at the H_0-constrained maximum likelihood estimate or constrained root of the likelihood estimate (CRLE) $\hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n)$ of the likelihood $q(\boldsymbol{y}_n; \boldsymbol{\xi}) = p(\boldsymbol{y}_n; \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{\xi}))$, and $\tilde{\boldsymbol{I}}_n^{-1}(\hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))$ is the n-sample Fisher information evaluated at the CMLE. In this context, the CMLE is the solution to the optimization problem $\max_{\boldsymbol{\xi}'} \log q(\boldsymbol{x}; \boldsymbol{\xi}')$ s.t. $\boldsymbol{h} \circ \boldsymbol{g}_{\boldsymbol{\theta}'}(\boldsymbol{\xi}') = \boldsymbol{0}$. As in theorem 3.5, $\tilde{\boldsymbol{s}}(\boldsymbol{y}_n; \boldsymbol{\xi}') = \boldsymbol{G}_{\boldsymbol{\theta}'}(\boldsymbol{\xi}')\boldsymbol{s}(\boldsymbol{y}_n; \boldsymbol{\theta}')$ and $\tilde{\boldsymbol{I}}_n(\boldsymbol{\xi}') = n\ \boldsymbol{G}^T(\boldsymbol{\xi}')\boldsymbol{I}(\boldsymbol{g}_{\boldsymbol{\theta}'}(\boldsymbol{\xi}'))\boldsymbol{G}(\boldsymbol{\xi}')$. Also, recall that for sufficiently large n, $\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n) = \boldsymbol{g}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\xi}}_{h(g)}(\boldsymbol{y}_n))$. Therefore, the locally (or asymptotically) equivalent Rao test statistic for the hypothesis in (3.39) is

$$\rho(\boldsymbol{y}_n) = \frac{1}{n}\boldsymbol{s}^T(\boldsymbol{y}_n; \hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\mathrm{CCRB}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{s}(\boldsymbol{y}_n; \hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n)), \tag{3.41}$$

which is analogous to (2.9) with the CCRB replacing the CRB. Under H_0, $\rho(\boldsymbol{y}_n)$ is still asymptotically χ^2_r in distribution. The corresponding Lagrange-multiplier variant of this statistic is given by

$$\rho(\boldsymbol{y}_n) = \frac{1}{n}\hat{\boldsymbol{\lambda}}_h(\boldsymbol{y}_n)\boldsymbol{H}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\mathrm{CCRB}(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}_h(\boldsymbol{y}_n))\hat{\boldsymbol{\lambda}}_h(\boldsymbol{y}_n), \tag{3.42}$$

where the Lagrange multiplier estimates $\hat{\boldsymbol{\lambda}}_h(\boldsymbol{y}_n)$ are based on the first order conditions relating to the constraint $\boldsymbol{h}$ (not $\boldsymbol{h}$ and $\boldsymbol{f}$).

The result in (3.41) is consistent with the classical results in [63], although not explicitly in this form. For the hypothesis scenario in (3.39) the Lagrange multiplier statistic should be $\hat{\boldsymbol{\lambda}}_h^T \boldsymbol{R}_{\hat{\boldsymbol{\theta}}}^{-1} \hat{\boldsymbol{\lambda}}_h$, where $\boldsymbol{R}_{\boldsymbol{h},\hat{\boldsymbol{\theta}}}$ is defined by

$$\begin{bmatrix} \boldsymbol{P}_{\boldsymbol{\theta}} & * & * \\ * & * & * \\ * & * & \boldsymbol{R}_{\boldsymbol{h},\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta}) & -\boldsymbol{F}^T(\boldsymbol{\theta}) & \boldsymbol{H}^T(\boldsymbol{\theta}) \\ \boldsymbol{F}(\boldsymbol{\theta}) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{H}(\boldsymbol{\theta}) & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}^{-1},$$

which is a variant of what appears in [63, equation (6.5)]. Finding the inverse using the Schur complement, it is clear that

$$\boldsymbol{R}_{\boldsymbol{h},\boldsymbol{\theta}} = \boldsymbol{H}^T(\boldsymbol{\theta}) \left[\boldsymbol{D}^{-1}(\boldsymbol{\theta}) - \boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) \left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta}) \right)^{-1} \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta}) \right] \boldsymbol{H}(\boldsymbol{\theta})$$

with $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})$. Recognizing the inner matrix as the Aitchison-Silvey-Crowder variant of the CCRB formula in (3.14) and substituting the CMLE for the parameter obtains (3.42).

3.5.2 The Wald statistic

Similarly, the Wald test statistic presented in section 2.5.2 is

$$\boldsymbol{h}^T(\boldsymbol{g}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))) \left(\boldsymbol{H}(\boldsymbol{g}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)))\boldsymbol{G}^T(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))\tilde{\boldsymbol{I}}_n^{-1}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))\boldsymbol{G}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))\boldsymbol{H}^T(\boldsymbol{g}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n))) \right)^{-1} \boldsymbol{h}(\boldsymbol{g}(\hat{\boldsymbol{\xi}}(\boldsymbol{y}_n)))$$

for the testing problem in (3.40), where $\boldsymbol{g}$ is localized about $\boldsymbol{\theta}$. Following the steps in section 3.5.1, then to test (3.39), the corresponding Wald test statistic is

$$\omega(\boldsymbol{y}_n) = n\boldsymbol{h}^T(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)) \left(\boldsymbol{H}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\mathrm{CCRB}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n))\boldsymbol{H}^T(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)) \right)^{-1} \boldsymbol{h}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)).$$

As with the Rao statistic, this general Wald statistic replaces the CRB in (2.10) with the CCRB, and under H_0, $\omega(\boldsymbol{y}_n)$ is asymptotically χ_r^2 in distribution.

This agrees with the classical result in [2, '$\lambda_{21}(\theta)$' on p. 240] where the Gorman-Hero-Aitchison-Silvey variant of the CCRB formula in (3.12) is used instead. (The general scenario when the FIM is singular is discussed in [2, section 3.9].)

A requirement for the existence of this statistic is that $\boldsymbol{H}(\boldsymbol{\theta})\mathrm{CCRB}(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta})$ be regular. This is not an additional requirement, but a necessity in testing that the hypothesis testing function itself be identifiable for the hypothesis to be valid.

3.6 Discussion

The previous sections have established that the theory of the constrained CRB is equivalent to that of the CRB. The majority of the proofs, besides their novelty in the recent literature on the CCRB, essentially rely on the generation of an implicit function that translates the constrained parametric problem into an unconstrained parametric problem, for which the theory of the CRB is well-established.

This theory has been extended to the CCRB, in particular, for identifiability under constraints; for the linear model with linear constraints (already well-documented in the literature but perhaps not with reference to this CCRB); for the constrained maximum likelihood problem, its asymptotic normality and the method of scoring; as well as for the Wald and Rao hypothesis tests. This list is by no means exhaustive. There exists recent research, for example, related to biased estimation with the CCRB [9], and an extension of the CCRB for complex-valued parameters [32]. In addition, there are other areas in mathematical statistics that, as far as I know, have not yet been connected to the CCRB, including a geometric interpreta-

tion of the CCRB (possibly in the manner of [60] or and extension of [5]), biasedness issues with constrained estimation, confidence intervals or sets, and the plausibility of a Bayesian version of the CCRB.

While the primary contributions of this thesis are the theoretical results and their proofs existing in this chapter, this should not discount the practical applications of these ideas. From the practitioners viewpoint, this research has produced a number of useful tools. For example, to test local identifiability, in addition to the classic result of Rothenberg (theorem 3.23), theorem 3.24 may be used. Likewise, for strict identifiability, theorem 3.29 is useful. And the method of scoring in (3.36) adds to the list of constrained maximum likelihood methods. These and others will be applied in a communications context, in the next chapter.

Chapter 4

Applications of the CCRB in Communications Models

Communications models in statistical signal processing often have the structure

$$y(n) = \mu(n, \boldsymbol{\theta}_n) + w(n) \ , \ n = 0, \ldots, N, \tag{4.1}$$

where $y(n)$ are the received observations of some model $\mu(n, \boldsymbol{\theta}_n)$ affected (additively) by noise $w(n)$ over a series of time samples n. The parameters may or may not be dependent on the number of time samples and the noise may or may not be independent (or even normally distributed). This general model encompasses a number of areas of signal processing, e.g., communications, sonar, radar, speech, imaging, control, sensing, networks, etc. In every one of these areas, there exist models where parametric equality constraints are of interest to practitioners in the field. The CCRB has proven useful as a performance analysis tool in localization [6], watermarking security [15], tomography [28], source bearing and symbol estimation [59], space-time block coding [42], and even a variant of least squares [46].

In this chapter, we shall detail just two rather general signal processing communications models: the convolutive channel and the calibrated array, each with unknown signal and channel components. Because the source and channel interact multiplicatively in the models, numerous variations of these two basic models are possible (and necessary). We shall formulate constraints for just a few of these vari-

ations and connect the results of this chapter to the theory developed in Chapter 3. For the sake of coherence in the presentation, the lengthier proofs of theorems in this chapter are relegated to appendix C.

In section 4.1, the convolutive mixture model with deterministic parameters is presented using a variety of descriptions. In addition, several useful terms are defined in section 4.1.1.2 that are useful to characterize conditions on the a notion of near-identifiability presented in section 4.1.2 and conditions on the Fisher information derived in 4.1.3. The corresponding complex-valued FIM (CFIM) is defined in 4.1.3.2 and important properties are given in 4.1.3.3. These properties are critical to understanding the particular inherent ambiguities in the basic convolutive mixture model and determining the class of constraints are necessary to obtain a CCRB, as discussed in 4.1.4.1. Two constraint models, the norm constraint (in 4.1.4.2) and the semiblind constraint (in 4.1.4.3) serve as further validation of the CCRB method by obtaining prior results in the literature; while another constraint model, the combination of the semiblind and unit modulus constraints in 4.1.4.4, demonstrates an important constraint model that did not previously exist in the literature.

In section 4.2, a special case of the convolutive mixture model, called the calibrated array model, is considered. The FIM for this (sub)model and its properties are detailed in section 4.2.1 and various constraints are considered in section 4.2.2. As before, a number of the constraint models are validations of the CCRB method compared with previous results in the literature (in 4.2.2.1, 4.2.2.2 and 4.2.2.5). The CCRB approach is also used to detail constraint models where the constraints are inappropriate either because the constraints are over-determined (in 4.2.2.3) or

because they are under-determined (in 4.2.2.4). As with the more general convolutive mixture model, a constraint model case is presented in 4.2.2.6 that did not previously exist in the literature.

4.1 Convolutive Mixture Model

The complex baseband representation of a multi-input, multi-output (MIMO) finite impulse response (FIR) system, or the convolutive mixture model, may be written as

$$\begin{aligned} y_m(n) &= \sum_{k=1}^{K} s^{(k)}(n) * h_m^{(k)}(n) + w_m(n) \\ &= \sum_{k=1}^{K} \sum_{l=0}^{L_k} s^{(k)}(n-l) h_m^{(k)}(l) + w_m(n) \end{aligned} \tag{4.2}$$

for the nth observation of the mth channel where the model consists of K sources, M channels ($1 \leq m \leq M$) with maximal order L_k for the kth source (or $L_k + 1$ taps for the kth source), N output samples ($0 \leq n \leq N-1$) per channel and $N + L_k$ input samples per channel per source. For each source, the model can be seen in figure 4.1. The dimensions K, M, L_k for $k = 1, \ldots, K$, and N are all assumed known. The elements $s^{(k)}(n)$ denote the scalar, complex value of the kth input source at time n; the elements of $h_m^{(k)}(l)$ denote the scalar, complex value of the lth filter coefficient (or lag) of the kth source processed by the mth channel. Both the signal inputs and channel coefficients are treated as deterministic unknowns, i.e., parameters having true values that can be estimated. The noise elements $w_m(n)$ are commonly assumed to be zero-mean, complex-valued, circular Gaussian iid over

space and time (and channel) with known variance σ^2.[1]

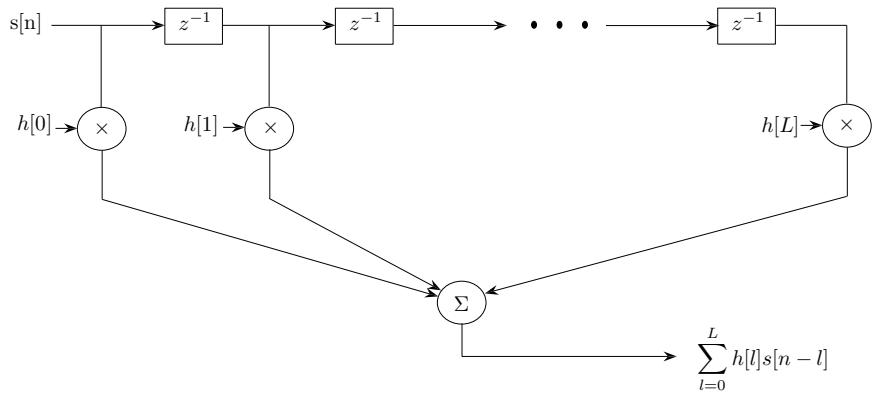


Figure 4.1: Finite Impulse Response (tapped delay line) model.

The convolution aspect of this model is often used to characterize the inter-symbol interference in direct sequence CDMA (code-division multiple-access) over dispersive channels [44]. This occurs, for example, when the propagation time of the signal is shorter than the coherence time of the channel in wideband signals. The varying channel order lengths L_k represent the different coherence times of the different source-channel links. This convolution also applies to scenarios where reverberation of the transmitted signal is present, which occurs when the communications is echoed. Additionally, the additive aspect over the sources represents the multiuser interference (e.g., the cocktail problem) common in communications systems today. Moreover, this convolutive mixture model incorporates a number of important model subclasses:

1. the convolutive single-input, multi-output (SIMO) model when $K = 1$,

[1]Under the Gaussian assumption, an unknown variance parameter decouples from the unknown mean parameters in the Fisher information. So, while the noise power will affect the performance potential, whether σ^2 is known or not does not affect how the CRB (CCRB) depends on the parameters. It does, of course, affect estimation.

2. the convolutive single-input, single-output (SISO) model when $K = M = 1$,

3. the memoryless, instantaneous mixing model when $L_k = 0$ for all $k = 1, \ldots, K$, and

4. the calibrated model (with constraints and a transformation of parameters).

For generality, we consider the full (MIMO) model, but the results contained in this section also apply to the preceding subclasses.

4.1.1 Equivalent Convolutive Mixture Models

4.1.1.1 The Vector-Matrix Model

In vector-matrix notation, the model may be written as

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{H}_M^{(k)} \boldsymbol{s}^{(k)} + \boldsymbol{w} = \boldsymbol{H}_M \boldsymbol{s} + \boldsymbol{w} \tag{4.3}$$

where the observations are contained in $\boldsymbol{y}^T = \left[\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_M^T\right] \in \mathbb{C}^{NM}$ and the observations for each channel contained in $\boldsymbol{y}_m^T = \left[y_m(0), \cdots, y_m(N-1)\right] \in \mathbb{C}^N$ for $m = 1, \ldots, M$, $s^{(k)T} = \left[s^{(k)}(-L_k), \ldots, s^{(k)}(N-1)\right] \in \mathbb{C}^{N+L_k}$ represents the input sequence of the kth source, the channel matrix is represented by the $NM \times N + L_k$ matrix

$$\boldsymbol{H}_M^{(k)} = \begin{bmatrix} \boldsymbol{H}_{(1)}^{(k)} \\ \vdots \\ \boldsymbol{H}_{(M)}^{(k)} \end{bmatrix} \tag{4.4}$$

where the mth channel submatrix over the kth source is given by the $N \times N + L_k$ matrix

$$\boldsymbol{H}_{(m)}^{(k)} = \begin{bmatrix} h_m^{(k)}(L_k) & h_m^{(k)}(L_k-1) & \cdots & h_m^{(k)}(0) & & \\ & h_m^{(k)}(L_k) & h_m^{(k)}(L_k-1) & \cdots & h_m^{(k)}(0) & \\ & & \ddots & & & \ddots \\ & & & h_m^{(k)}(L_k) & h_m^{(k)}(L_k-1) & \cdots & h_m^{(k)}(0) \end{bmatrix} \tag{4.5}$$

for $m = 1, \ldots, M$, and the noise vector is $\boldsymbol{w}^T = \left[\boldsymbol{w}_1^T, \cdots, \boldsymbol{w}_M^T\right] \in \mathbb{C}^{MN}$ with the noise for each channel given by $\boldsymbol{w}_m^T = \left[w_m(0), \ldots, w_m(N-1)\right] \in \mathbb{C}^N$. This particular vector-matrix ordering of the model can also be represented as

$$\boldsymbol{y} = \sum_{k=1}^{K} \left(\boldsymbol{I}_{M\times M} \otimes \boldsymbol{S}^{(k)}\right) \boldsymbol{h}^{(k)} + \boldsymbol{w} = \boldsymbol{S}_M \boldsymbol{h} + \boldsymbol{w} \tag{4.6}$$

where the input samples are now organized into an $N \times L_k + 1$ Toeplitz matrix

$$\boldsymbol{S}^{(k)} = \begin{bmatrix} s^{(k)}(0) & s^{(k)}(-1) & \cdots & s^{(k)}(-L_k) \\ s^{(k)}(1) & s^{(k)}(0) & \cdots & s^{(k)}(-L_k+1) \\ \vdots & \vdots & \ddots & \vdots \\ s^{(k)}(N-1) & s^{(k)}(N-2) & \cdots & s^{(k)}(N-L_k-1) \end{bmatrix}$$

with $\otimes$ being the Kronecker product, and the channel elements are vectorized as $\boldsymbol{h}^{(k)T} = \left[\boldsymbol{h}_1^{(k)T}, \ldots, \boldsymbol{h}_M^{(k)T}\right] \in \mathbb{C}^{M(L_k+1)}$ with $\boldsymbol{h}_m^{(k)T} = \left[h_m^{(k)}(0), \ldots, h_m^{(k)}(L_k)\right] \in \mathbb{C}^{L_k+1}$ for each $k = 1, \ldots, K$ and $m = 1, \ldots, M$. The purpose of these alternative vector-matrix methods will become evident in the development of the Fisher information (see section 4.1.3) for the original model in (4.2).

4.1.1.2 The Z transform model

Yet another alternative representation of this model, necessary to introduce relevant characteristics of the model parameters, and often referred to as being in

the Z-transform domain, is the M-variate stationary process

$$\boldsymbol{y}(n) = \big[\boldsymbol{H}(z)\big] * \boldsymbol{s}(n) + \boldsymbol{w}(n) \tag{4.7}$$

where $\boldsymbol{H}(z)$ is a $M \times K$ (global) transfer function (polynomial) defined by

$$\boldsymbol{H}(z) = \big[\boldsymbol{H}^{(1)}(z), \ldots, \boldsymbol{H}^{(K)}(z)\big]$$

for nonzero $z \in \mathbb{C}^*$ (including ∞) with $\boldsymbol{H}^{(k)}(z)$ being the kth source transfer function

$$\boldsymbol{H}^{(k)}(z) = \sum_{l=0}^{L_k} \boldsymbol{h}^{(k)}(l) z^{-l}$$

and with $\boldsymbol{h}^{(k)}(l)^T = \Big[h_1^{(k)}(l), h_2^{(k)}(l), \ldots, h_M^{(k)}(l)\Big] \in \mathbb{C}^M$ for $l = 0, \ldots, L_k$ and $k = 1, \ldots, K$, $\boldsymbol{s}(n)^T = \big[s^{(1)}(n), s^{(2)}(n), \ldots, s^{(K)}(n)\big] \in \mathbb{C}^K$ for $n = 0, \ldots, N-1$, and $\boldsymbol{y}(n)$ and $\boldsymbol{w}(n)$ correspond to match the models in (4.3) and (4.6).

Hence, $\boldsymbol{H}(z)$ is a polynomial matrix of the backward shift z^{-1}. The kth source transfer function is said to have a *common zero* if there exists a nonzero $z_0 \in \mathbb{C}^*$ such that $\boldsymbol{H}^{(k)}(z_0) = \boldsymbol{0}$. If the transfer function does not have a common zero, then the polynomials $\boldsymbol{H}_{(m)}^{(k)}(z) = \sum_{l=0}^{L_k} h_m^{(k)}(l) z^{-l}$, for $m = 1, \ldots, M$, are said to be *coprime*. The global transfer function is said to be *reducible* if there exists a nonzero $z_0 \in \mathbb{C}^*$ such that $\operatorname{rank}(\boldsymbol{H}(z_0)) < K$. If not, it is said to be *irreducible*. The global transfer function is said to be *column-reduced* if $\lim_{z \to 0} \operatorname{rank}(\boldsymbol{H}(z)\boldsymbol{Z}) = K$ where $\boldsymbol{Z} = \operatorname{diag}\{z^{L_1}, \ldots, z^{L_K}\}$. This necessarily implies for an irreducible and column-reduced transfer function, called *a minimum polynomial basis* [35], that $M \geq K$. The connection between irreducibility in the multi-source case and not having common zeros in the single source case is made clear in the following result.

Theorem 4.1. $\boldsymbol{H}(z)$ is irreducible if and only if $\sum_{k=1}^{K} \alpha_k \boldsymbol{H}^{(k)}(z)$ has no common zeros for any nontrivial complex-valued collection $\alpha_1, \ldots, \alpha_K$.

Proof. $\boldsymbol{H}(z)$ is reducible if and only if there exists some point $z_0 \in \mathbb{C}^*$ such that $\text{rank}(\boldsymbol{H}(z_0)) < K$, which follows if and only if there exists nontrivial $\beta_1, \ldots, \beta_K \in \mathbb{C}$ such that $\sum_{k=1}^{K} \beta_k \boldsymbol{H}^{(k)}(z_0) = \boldsymbol{0}$, i.e., if and only if $\sum_{k=1}^{K} \beta_k \boldsymbol{H}^{(k)}(z)$ has a common zero. □

The connection between the model in (4.3) and (4.7) is that the K-source, M-channel matrix $\boldsymbol{H}_M$ is a column-rotated, generalized Sylvester matrix of the block Toeplitz matrix

$$\begin{bmatrix} \boldsymbol{H}(L) & \boldsymbol{H}(L-1) & \cdots & \boldsymbol{H}(0) & & \\ & \boldsymbol{H}(L) & \boldsymbol{H}(L-1) & \cdots & \boldsymbol{H}(0) & \\ & & \ddots & & & \ddots \\ & & & \boldsymbol{H}(L) & \boldsymbol{H}(L-1) & \cdots & \boldsymbol{H}(0) \end{bmatrix}$$

where $\boldsymbol{H}(l) = \left[\boldsymbol{h}^{(1)}(l), \ldots, \boldsymbol{h}^{(K)}(l)\right]$ with $\boldsymbol{h}^{(k)}(l)$ a null vector when $l > L_k$ and $L = \max_k L_k$. Furthermore, under certain conditions, the rank of $\boldsymbol{H}_M$ is determined by the characterization of the basis $\boldsymbol{H}(z)$.

Theorem 4.2. Assume $M > K$ and $N \geq \sum_{k=1}^{K} L_k$. Then $\boldsymbol{H}_M$ is full column rank if and only if $\boldsymbol{H}(z)$ is a minimum polynomial basis.

Proof. This can be shown by a variant of the proof in Loubaton and Moulines [44, theorem 1]. □

The input sequence $\boldsymbol{s}^{(k)}$ is said to have p_k *modes*[2] if it can be written as a

[2]There are a number of alternative definition of modes [30, 43], [35, p. 168].

linear combination

$$s^{(k)}(n) = \sum_{i=1}^{p_k} c_i m_i^{n+L_k},$$

where c_i, $i = 1, \ldots, p_k$ are complex-valued weights and m_i, $i = 1, \ldots, p_k$ are the $\mathbb{C}^*$-valued roots of the polynomial

$$a(0) + a(1)z^{-1} + a(2)z^{-2} + \cdots + a(p_k)z^{-p_k},$$

whose coefficients satisfy

$$\sum_{j=0}^{p_k} s^{(k)}(i+j)a(j) = 0$$

for $i = -L_k, \ldots, N - p_k - 1$. Hence, the Toeplitz matrix

$$\boldsymbol{S}^{(k)}(n) = \begin{bmatrix} s^{(k)}(n) & s^{(k)}(n-1) & \cdots & s^{(k)}(-L_k) \\ s^{(k)}(n+1) & s^{(k)}(n) & \cdots & s^{(k)}(-L_k+1) \\ \vdots & & \ddots & \vdots \\ s^{(k)}(N-1) & s^{(k)}(N-2) & \cdots & s^{(k)}(N-n-L_k-1) \end{bmatrix} \tag{4.8}$$

has rank $\min\{N - n, p, n + L_k + 1\}$ [76, lemma 1]. Note, $\boldsymbol{S}^{(k)}(0) = \boldsymbol{S}^{(k)}$ and $\boldsymbol{S}^{(k)}(-L_k) = \boldsymbol{s}^{(k)}$ from section 4.1.1.1. If one of the modes of $\boldsymbol{s}^{(k)}$ is a common zero of the channel transfer function $\boldsymbol{H}^{(k)}(z)$, say m_v, then the channel makes no distinction between $\boldsymbol{s}^{(k)}$ and $\boldsymbol{s}^{(k)'}$, defined by

$$s^{(k)'}(n) = \sum_{\substack{i=1 \\ i \neq v}}^{p_k} c_i m_i^{n+L_k},$$

even though $\boldsymbol{s}^{(k)} \neq \boldsymbol{s}^{(k)'}$. Ironically, if a channel lacks sufficient diversity, the more modes an input has leads to greater risk of lost information on the input but potentially less meaningful loss depending on the weights.

Theorem 4.3. The matrix

$$\boldsymbol{S}(n) = \begin{bmatrix} \boldsymbol{S}^{(1)}(n) & \boldsymbol{S}^{(2)}(n) & \cdots & \boldsymbol{S}^{(K)}(n) \end{bmatrix} \tag{4.9}$$

is full column rank only if $N \geq (K+1)n + K + \sum_{j=1}^{K} L_j$, $p_{\text{total}} \geq Kn + K + \sum_{k=1}^{K} L_k$, and $p_k \geq L_k + 1 + n$ for each $k = 1, \ldots, K$ and if $N \geq (K+1)n + K + (K+2)\sum_{j=1}^{K} L_j$, $p_{\text{total}} \geq K * n + K + (K+1)\sum_{k=1}^{K} L_k$, and $p_k \geq L_k + 1 + n + \sum_{j=1}^{K} L_j$.

Proof. This is a variation of the results in [76, 1, 48, 49]. □

In particular, for $n = 0$ then $\boldsymbol{S}(0)$ requires $N \geq K + \sum_{j=1}^{K} L_j$, $p_{\text{total}} \geq K + \sum_{j=1}^{K} L_j$, and $p_k \geq L_k + 1$, for each $k = 1, \ldots, K$, to be full rank. Or conversely, $\boldsymbol{S}(0)$ is full rank if $N \geq K + (K+2)\sum_{j=1}^{K} L_j$, $p_{\text{total}} \geq K + (K+1)\sum_{k=1}^{K} L_k$, and $p_k \geq L_k + 1 + \sum_{j=1}^{K} L_j$.

Before continuing to the development of the Fisher information for this convolutive mixture model, it is relevant to consider a notion of identifiability using the concepts of the Z-transform model in (4.7). However, this next section is a brief aside that may be skipped if not of interest to the reader.

4.1.2 Strict Identifiability

The K-user, M-channel FIR system in (4.2) or (4.3) is said to be *strictly identifiable (SI)* if and only if

$$\boldsymbol{H}_M \boldsymbol{s} = \boldsymbol{H}_M' \boldsymbol{s}' \iff \boldsymbol{H}'(z) = \boldsymbol{H}(z)\boldsymbol{A} \text{ and } \boldsymbol{s}'(n) = \boldsymbol{A}^{-1}\boldsymbol{s}(n)$$

for some nonsingular matrix $\boldsymbol{A}$. The converse statement is always true, so strict identifiability depends on the conditional statement. The term strict identifiability is a misnomer. As is clear from the definition, the deterministic parameters are not "identifiable" when they are strictly identifiable. Instead, the parameters are

identifiable up to some minimal ambiguity. When this situation occurs, then it is possible to treat the channel (signal) components statistically to truly identify the deterministic signal (channel) parameters using stochastic approaches, e.g., the subspace method [24, 44]. In the context of this thesis, where the parameters are not treated as random but as deterministic, the reduction to a minimal ambiguity set also reduces the number of necessary constraints to eliminate it.

Strict identifiability in the convolutive channel model, it shall be shown, has some interesting connections with the Fisher information matrix. This is not surprising, given the results in section 3.2. It is perhaps also intuitive to expect that as strict identifiability is a notion of *near-identifiability*, then the corresponding Fisher information will satisfy some notion of *near-regularity*.

Abed-Meraim and Hua [1] developed necessary and sufficient conditions for strict identifiability in terms of characteristics in the Z-transform model, i.e., no channel zeros, the number of signal modes, more sources than sensors, etc. The following two theorems will be presented without proof, as they were in [1]. Proofs I developed can be found in [48].

Theorem 4.4 (SI necessary conditions)**.** The M-channel K-source FIR system is strictly identifiable only if[3]

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + \sum_{j=1}^{K} L_j$,

[3]In [1], condition (a) omits the column-reducedness requirement, condition (c) did not include the special case, and condition (d) is originally (and I think erroneously) stated $N \geq 2K + \sum_{j=1}^{K} L_j$.

(c) $p_k \geq L_k + 2$ for $k = 1, \ldots, K$ of $p_k \geq 1$ if $L_k = 0$, and

(d) $N \geq K + \sum_{j=1}^{K} L_j$.

Theorem 4.5 (SI sufficiency conditions)**.** The M-channel K-source FIR system is strictly identifiable if

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + (K+1)\sum_{j=1}^{K} L_j$,

(c) $p_k \geq L_k + 1 + \sum_{j=1}^{K} L_j$ for $k = 1, \ldots, K$, and

(d) $N \geq K + (K+2)\sum_{j=1}^{K} L_j$.

Yet other notions of near-identifiability in the convolutive channel model exist in the literature, e.g., *cross-relation-based identifiability* [43], which have an equivalence to strict identifiability in the SIMO case [30]. The necessary and sufficient conditions presented here are independent of the number of channels, although one should expect that increasing the number of channels would increase channel diversity thereby weakening the requirements on the other channel or source characteristics. Hence, theorems 4.4 and 4.5 are in some sense conditions for the least diversified, strictly identifiable scenario $M = K + 1$. Table 4.1 shows the growth in necessary and sufficient conditions as the number of users increases for a fixed channel size and for fixed channel orders.

Table 4.1: Necessary and sufficient data sizes for strict identifiability

# of sources K	1	2	3	4	5
necessary data size N	6	12	18	24	30
sufficient data size N	16	42	78	124	180

($M = 6$ channels, $L_k = 5$ for all sources)

4.1.3 The Fisher information of the convolutive mixture model

4.1.3.1 Complex-valued Fisher information

Before presenting the Fisher information on the model in (4.2), certain details about Fisher information matrices on complex-valued parameters are necessary. The structure of the FIM depends on the ordering of the parameters. For complex-valued parameters, the FIM parameter vector consists of the real and imaginary parts of the parameters. If the complex-valued parameters are collected in the vector $\boldsymbol{\vartheta}$ and the (FIM) real parameter vector is defined to be $\boldsymbol{\theta}^T = \left[\mathrm{Re}(\boldsymbol{\vartheta}^T), \mathrm{Im}(\boldsymbol{\vartheta}^T)\right]$ and the real-valued parameter FIM has the structure

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{E}_1(\boldsymbol{\theta}) & -\boldsymbol{E}_2(\boldsymbol{\theta}) \\ \boldsymbol{E}_2(\boldsymbol{\theta}) & \boldsymbol{E}_1(\boldsymbol{\theta}) \end{bmatrix}, \tag{4.10}$$

then the *complex-valued parameter Fisher information matrix* (CFIM) may be defined as[4]

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}) = E_{\boldsymbol{\vartheta}} \boldsymbol{\jmath}(\boldsymbol{x}; \boldsymbol{\vartheta}) \boldsymbol{\jmath}^H(\boldsymbol{x}; \boldsymbol{\vartheta}) = \frac{1}{2} \left(\boldsymbol{E}_1(\boldsymbol{\theta}) + j \cdot \boldsymbol{E}_2(\boldsymbol{\theta}) \right)$$

where $\boldsymbol{\jmath}(\boldsymbol{x}; \boldsymbol{\vartheta}) = \dfrac{\partial \log p(\boldsymbol{x}; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*}$.[5] A number of properties for the real-valued parameter FIM can be gleaned from properties on the CFIM.

[4] This CFIM is a submatrix of the complex-valued parameter FIM developed by van den Bos [70], which would be the preferred FIM for use in a performance metric or in applying constraints [32]. However, in this and the next sections, the CFIM presented here is only used to obtain properties relevant to the real-valued parameter FIM.

[5] The complex derivative is defined to be $\frac{\partial}{\partial \boldsymbol{\vartheta}} = \frac{\partial}{\partial \mathrm{Re}\boldsymbol{\vartheta}} - j \cdot \frac{\partial}{\partial \mathrm{Im}\boldsymbol{\vartheta}}$ in this thesis.

Theorem 4.6. The null space of the FIM of the form (4.10) has dimension exactly twice the dimension of the null space of the corresponding CFIM, i.e.,

$$\text{nullity}(\boldsymbol{I}(\boldsymbol{\theta})) = 2 \cdot \text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})).$$

Proof. Note $\begin{bmatrix} \boldsymbol{E}_1 & -\boldsymbol{E}_2 \\ \boldsymbol{E}_2 & \boldsymbol{E}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \boldsymbol{0}$ if and only if $\boldsymbol{E}_1\boldsymbol{a} - \boldsymbol{E}_2\boldsymbol{b} = \boldsymbol{0}$ and $\boldsymbol{E}_2\boldsymbol{a} + \boldsymbol{E}_1\boldsymbol{b} = \boldsymbol{0}$ if and only if $\left(\boldsymbol{E}_1 + j\boldsymbol{E}_2\right)\left(\boldsymbol{a} + j\boldsymbol{b}\right) = \left(\boldsymbol{E}_1\boldsymbol{a} - \boldsymbol{E}_2\boldsymbol{b}\right) + j\left(\boldsymbol{E}_2\boldsymbol{a} + \boldsymbol{E}_1\boldsymbol{b}\right) = \boldsymbol{0}$. Also, $\begin{bmatrix} \boldsymbol{E}_1 & -\boldsymbol{E}_2 \\ \boldsymbol{E}_2 & \boldsymbol{E}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \boldsymbol{0}$ if and only if $\begin{bmatrix} \boldsymbol{E}_1 & -\boldsymbol{E}_2 \\ \boldsymbol{E}_2 & \boldsymbol{E}_1 \end{bmatrix} \begin{bmatrix} -\boldsymbol{b} \\ \boldsymbol{a} \end{bmatrix} = \boldsymbol{0}$. Finally, the dimension of $\begin{bmatrix} \boldsymbol{a} & -\boldsymbol{b} \\ \boldsymbol{b} & \boldsymbol{a} \end{bmatrix}$ is 2 unless $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{0}$; however, the dimension of $\begin{bmatrix} \boldsymbol{a} + j\boldsymbol{b}, & -\boldsymbol{b} + j\boldsymbol{a} \end{bmatrix}$ is only 1 since $\left(-\boldsymbol{b} + j\boldsymbol{a}\right) = -j \cdot \left(\boldsymbol{a} + j\boldsymbol{b}\right)$. □

A version of Bang's formula [36, a variant of (15.52) on p. 525] can be developed for complex-valued parameters, i.e.,

$$\begin{aligned} \boldsymbol{\mathcal{I}}_{ij}(\boldsymbol{\vartheta}) \;\; = \;\; & \text{tr}\left[\boldsymbol{C}^{-1}(\boldsymbol{\vartheta})\frac{\partial \boldsymbol{C}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_i^*}\boldsymbol{C}^{-1}(\boldsymbol{\vartheta})\frac{\partial \boldsymbol{C}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_j}\right] \\ & + \frac{\partial \boldsymbol{\mu}^H(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_i^*}\boldsymbol{C}^{-1}(\boldsymbol{\vartheta})\frac{\partial \boldsymbol{\mu}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_j} + \frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_i^*}\boldsymbol{C}^{-1}(\boldsymbol{\vartheta})\frac{\partial \boldsymbol{\mu}^*(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_j} \end{aligned}$$

for any observations $\boldsymbol{y} \sim \mathcal{CN}(\boldsymbol{\mu}(\boldsymbol{\vartheta}), \boldsymbol{C}(\boldsymbol{\vartheta}))$.

4.1.3.2 CFIM for the convolutive mixture model

For the convolutive mixture model in (4.3) and (4.6), only the mean vector depends on the unknown parameters, hence

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}) = \frac{\partial \boldsymbol{\mu}^H(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{\mu}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^T} + \frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{\mu}^*(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^T}.$$

The mean vector is

$$\boldsymbol{\mu}(\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \boldsymbol{H}_M^{(k)}\boldsymbol{s}^{(k)} = \sum_{k=1}^{K} \left(\boldsymbol{I}_M \otimes \boldsymbol{S}^{(k)}\right)\boldsymbol{h}^{(k)}.$$

Therefore, if the complex-valued parameter vector in the model in (4.3) is defined by

$$\boldsymbol{\vartheta}^T = \begin{bmatrix} \boldsymbol{h}^{(1)T} \\ \boldsymbol{s}^{(1)T} \\ \vdots \\ \boldsymbol{h}^{(K)T} \\ \boldsymbol{s}^{(K)T} \end{bmatrix} \tag{4.11}$$

then the complex-valued FIM of the model can be shown to be

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}) = \frac{1}{\sigma^2}\boldsymbol{\mathcal{Q}}^H\boldsymbol{\mathcal{Q}} = \frac{1}{\sigma^2}\begin{bmatrix} \boldsymbol{Q}_1^H\boldsymbol{Q}_1 & \boldsymbol{Q}_1^H\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_1^H\boldsymbol{Q}_K \\ \boldsymbol{Q}_2^H\boldsymbol{Q}_1 & \boldsymbol{Q}_2^H\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_2^H\boldsymbol{Q}_K \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{Q}_K^H\boldsymbol{Q}_1 & \boldsymbol{Q}_K^H\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_K^H\boldsymbol{Q}_K \end{bmatrix} \tag{4.12}$$

where $\boldsymbol{\mathcal{Q}} = \begin{bmatrix}\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K\end{bmatrix}$ and $\boldsymbol{Q}_k = \begin{bmatrix}\boldsymbol{I}_M \otimes \boldsymbol{S}^{(k)} & \boldsymbol{H}_M^{(k)}\end{bmatrix}$.

4.1.3.3 Properties of the CFIM

In this section, I develop properties of this CFIM. In particular, the singularity of the CFIM is proven and a limit on the dimension of this singularity is detailed, as well as necessary and sufficient conditions on the parameter characteristics to attain this limit.

Given the inherent relationship between regularity of the FIM and identifiability as detailed in section 3.2, it should not be surprising that the FIM, and hence CFIM, is singular. The model presented in (4.3) or (4.6) has a multiplicative ambiguity with any source interacting with its corresponding channel, i.e., for any nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{L_k+1 \times L_k+1}$, the input source matrix $\boldsymbol{S}^{(k)}$ and the channel vector $\boldsymbol{h}_m^{(k)}$ are indistinguishable from $\boldsymbol{S}^{(k)}\boldsymbol{A}$ and $\boldsymbol{A}^{-1}\boldsymbol{h}_m^{(k)}$, respectively. Over all the channels, this presumes a complex-valued multiplicative ambiguity of at least ML_k. Additionally, cross-source ambiguities exist between a source input and a different

source channel. The limit for the minimal degrees of freedom or the rank of the ambiguity is given in the following theorem.

Theorem 4.7. The CFIM is singular and the dimension of its null space is lower bounded as

$$\text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) \geq \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+, \tag{4.13}$$

where $(a)_+ = a$ for $a \geq 0$ and $(a)_+ = 0$ for $a < 0$. This limit quantity is the *nullity lower bound (NLB).*

The proof can be found in the appendix C. The proof constructs the following matrix, whose linearly independent columns are a basis for the null space of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$,

$$\boldsymbol{\mathcal{N}} = \begin{bmatrix} \boldsymbol{h}^{(1)} & \boldsymbol{\mathcal{H}}^{(2)}_{(1)} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\mathcal{H}}^{(K)}_{(1)} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ -\boldsymbol{s}^{(1)} & \mathbf{0} & \mathbf{0} & -\boldsymbol{\mathcal{S}}^{(2)}_{(1)} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & -\boldsymbol{\mathcal{S}}^{(K)}_{(1)} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{h}^{(2)} & \boldsymbol{\mathcal{H}}^{(1)}_{(2)} & \cdots & \mathbf{0} & \boldsymbol{\mathcal{H}}^{(K)}_{(2)} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\mathcal{S}}^{(1)}_{(2)} & -\boldsymbol{s}^{(2)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & -\boldsymbol{\mathcal{S}}^{(K)}_{(2)} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{h}^{(K)} & \cdots & \boldsymbol{\mathcal{H}}^{(2)}_{(K)} & \boldsymbol{\mathcal{H}}^{(1)}_{(K)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & -\boldsymbol{\mathcal{S}}^{(1)}_{(K)} & -\boldsymbol{\mathcal{S}}^{(2)}_{(K)} & \cdots & -\boldsymbol{s}^{(K)} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{4.14}$$

where $\boldsymbol{\mathcal{H}}^{(i)}_{(j)}$ and $\boldsymbol{\mathcal{S}}^{(i)}_{(j)}$ are defined in (C.2) and (C.4), respectively.

The nullity lower bound (NLB) seems to be an unusual quantity. The term $(L_i - L_j + 1)_+$ represents the ambiguity for the "tap window" of the ith channel masking the interaction of the jth channel with its corresponding source. This quantity also reveals that an ambiguity exists only if the coherence-propagation delay of the ith channel is at least as greats as for the jth channel. Intuitively, the greater the spread between the channel orders increases the overlapping window for the jth channel's interaction to be masked and thereby increases the ambiguity. Conversely, when the channel orders have narrow differences, this increases the

diversity of the channel by limiting the window where one channel can cover that of another.

A simple corollary follows stating that the NLB or the dimension of the ambiguity set is at least the square of the number of sources.

Corollary 4.8. nullity($\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$) $\geq K^2$.

Proof. Note $(L_i - L_j + 1)_+ + (L_j - L_i + 1)_+ \geq 2$ with equality if and only if $|L_i - L_j| \leq 1$. Thus,

$$\begin{aligned}
\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+ &= \sum_{i=1}^{K}\sum_{\substack{j=1\\ j\neq 1}}^{K}(L_i - L_j + 1)_+ + \sum_{i=1}^{K}(L_i - L_i + 1)_+ \\
&= \sum_{i=1}^{K}\sum_{j=i+1}^{K}(L_i - L_j + 1)_+ + (L_j - L_i + 1)_+ + K \\
&\geq \frac{K(K-1)}{2}\cdot 2 + K = K^2.
\end{aligned}$$

□

The proof also reveals that this K^2 degree of uncertainty can only be attained if any two orders differ by at most 1 tap, which agrees with the intuition that channel diversity is enhanced when the channel orders are not widely spread.

In addition to the ambiguity due to the channel order spread, the degrees of freedom in the model depends on the number of parameters and the number of observations.

Theorem 4.9. nullity($\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$) $\geq \sum_{k=1}^{K}(N + L_k + M(L_k + 1)) - MN$.

Proof. Since $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}) = \frac{1}{\sigma^2}\boldsymbol{\mathcal{Q}}^H\boldsymbol{\mathcal{Q}}$ then rank($\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$) = rank($\boldsymbol{\mathcal{Q}}$). Since $\boldsymbol{\mathcal{Q}}$ is a $MN \times \sum_{k=1}^{K}(N+L_k+M(L_k+1))$, then $\text{rank}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) \leq \min\left\{MN, \sum_{k=1}^{K}(N + L_k + M(L_k + 1))\right\}$

and therefore

$$\begin{aligned} \text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) &\geq \text{colsize}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) - \text{rank}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) \\ &\geq \max\left\{0\ ,\ \sum_{k=1}^{K}(N + L_k + M(L_k+1)) - MN\right\}. \end{aligned}$$

□

The number of columns of $\boldsymbol{\mathcal{Q}}$ (or $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$) corresponds to the number of unknown parameters. Likewise, the number of rows of $\boldsymbol{\mathcal{Q}}$ corresponds to the number of equations (observations) in the model (4.2). In most scenarios, having more equations (rows) than unknowns (columns) is a necessary requirement for the unknowns to be solvable. Thus, the degrees of freedom of the ambiguity space is at least the number of unknowns (parameters) less the number of equations. However, if the equations are linearly dependent (redundant), as is the case in the convolutive mixture model, then the degrees of freedom is potentially greater. The following corollary combines theorems 4.7 and 4.9.

Corollary 4.10.

$$\text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) \geq \max\left\{\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+, \sum_{k=1}^{K}(N + L_k + M(L_k+1)) - MN\right\}.$$

The NLB depends only on the channel orders and number of sources, whereas the "equations vs. unknowns" nullity bound depends on the channel orders, the sample size per source, the number of subchannels per source, and the number of sources. Hence, it is of interest to determine under what conditions on the sample size N and the number of channels M per source will this second bound be of no consequence. Control over the dimensions K (the number of transmitters), M

(the number of receivers), N (the number of time snapshots or transmission length) is possible in the design of many communications models. The next two results determine conditions on the dimensions of the model which allow the NLB to be the minimum possible degrees of freedom. The first condition requires more subchannels per source than sources (or more receivers than transmitters in a communications context).

Theorem 4.11. The CFIM $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ can attain the CFIM nullity lower bound only if $M > K$.

Proof. If $M \leq K$ then $\boldsymbol{H}_M$, which is a $MN \times KN + \sum_{k=1}^{K} L_k$ matrix, cannot be full column rank by theorem 4.2. This implies the existence of a null space of a larger dimension than the space spanned by the columns of $\boldsymbol{\mathcal{N}}$ in (4.14). Hence $\text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) > \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+$. □

Under the assumption that more receivers than transmitters are in the model, then corollary 4.10 implies a minimal requirement on the transmission snapshots.

Theorem 4.12. Provided $M > K$, the CFIM $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ can attain the CFIM nullity lower bound only if

$$N \geq \frac{K+1}{M-K}\sum_{j=1}^{K} L_j + \sum_{j=1}^{K} L_j + K + \frac{1}{M-K}\left(K^2 - \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+\right) \quad (4.15)$$

Proof. We desire $\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+ \geq \sum_{k=1}^{K}(N + L_k + M(L_k + 1)) - MN$. Solving for N shows the result. □

As the channel diversity (in terms of the number of channels per source) increases, the necessary size on the data to attain the CFIM NLB decreases. As

$M \to \infty$ then the bound on the sample size becomes simply $N \geq K + \sum_{j=1}^{K} L_j$, which is comparable to theorem 4.4(d). However, for any finite M and nontrivial channel orders, the bound is effectively $N \geq K + \sum_{j=1}^{K} L_j + 1$ since $-\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+ \geq -K\sum_{j=1}^{K}(L_j + 1)$, and hence a looser bound than (4.15) is

$$\begin{aligned} N &\geq \frac{K+1}{M-K}\sum_{j=1}^{K} L_j + \sum_{j=1}^{K} L_j + K + \frac{1}{M-K}\left(K^2 - K\sum_{j=1}^{K}(L_j+1)\right) \\ &\geq \sum_{j=1}^{K} L_j + K + \frac{1}{M-K}\sum_{j=1}^{K} L_j. \end{aligned}$$

In the scenario requiring the most data samples, $M = K + 1$, then the system requires $N \geq (K+1)\sum_{j=1}^{K} L_j + \sum_{j=1}^{K} L_j + K + \left(K^2 - \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+\right)$. (This last term is always nonpositive.) In the SIMO scenario ($K = 1$), this necessary condition becomes $N \geq 3L + 1$ which agrees with the requirement in [31].[6]

Given that $M > K$ and N satisfies (4.15), then the minimum possible nullity of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ is the nullity lower bound,

$$\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+,$$

i.e., it is possible to limit the ambiguity strict to the mixing of sources in the convolutive channel. If it is possible to understand the conditions under which the CFIM attains this nullity lower bound, then it is known that the null space is completely characterized by the columns of $\boldsymbol{\mathcal{N}}$ in (4.14). The importance of this is that when

[6]For the SIMO ($K = 1$) case, this nullity lower bound is exactly 1 and parameters for which the CFIM attains this bound are said to be *Fisher information identifiable* in [30]. The notion of identifiability for the SIMO case is sensible since by scaling a single parameter it is possible to obtain a bound on all the remaining parameters relative to the scaled parameter. This naturally connects with the notion of strict identifiability in section 4.1.2. This interpretation does not extend simply to the MIMO ($K > 1$) case; however, it shall be shown that there does exist an inherent connection between a CFIM attaining the NLB and the notion of strict identifiability.

the null space can be parameterized, then it is possible to use theorems 3.23 and 3.24 to determine constraints that lead to regularity of the CFIM (and FIM). Using the concepts of signal excitation (modes) and channel diversity (irreducibility and column-reducedness) as defined in section 4.1.1.2, the following theorems also establish a correlation of conditions between the idea of near-regularity when the FIM attains the NLB and of near-identifiability (or strict identifiability) of section 4.1.2.

Theorem 4.13 (CFIM NLB necessary conditions)**.** The M-channel K-source FIR system Fisher information matrix has a nullity of exactly the NLB in (4.13) only if

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + \sum_{j=1}^{K} L_j$,

(c) $p_k \geq L_k + 2$ for $k = 1, \ldots, K$ or $p_k \geq 1$ if $L_k = 0$,

(d) $N \geq K + \sum_{j=1}^{K} L_j$, and

(e) $M > K$.

Theorem 4.14 (CFIM NLB sufficiency conditions)**.** The M-channel K-source FIR system FIM has a nullity of exactly the NLB in (4.13) if

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + (K+1) \sum_{j=1}^{K} L_j$,

(c) $p_k \geq L_k + 1 + \sum_{j=1}^{K} L_j$ for $k = 1, \ldots, K$,

Table 4.2: Necessary and sufficient data sizes for the FIM to attain the NLB

# of sources K	1	2	3	4	5
necessary data size N	8	20	38	74	180
sufficient data size N	16	42	78	124	180

($M = 6$ channels, $L_k = 5$ for all sources)

(d) $N \geq K + (K+2)\sum_{j=1}^{K} L_j$, and

(e) $M > K$.

With the exception of condition (e), then theorems 4.13 and 4.14 are identical to theorems 4.4 and 4.5, respectively. This is an expected result since the Fisher information for identifiable parametric components within the model should meet certain regularity conditions (see theorem 2.5). For theorem 4.13(d), the condition is weaker than the condition in theorem 4.12, but it is left weaker to agree with the necessary condition for strict identifiability in theorem 4.4(d). (It is unclear if that condition would change if dependence on channel size M is considered.) Table 4.1.3.3 show the growth in the necessary and sufficient data size N as the number of sources increases for fixed channel size and channel orders, using the stronger condition in theorem 4.12 instead of that in theorem 4.13(d). It is clear that in comparison, as the number of sources increases the number of channels, the necessary data size grows to match the sufficient data size, which lead to theorem 4.15 and corollary 4.16.

First, the special case when the CFIM attains the minimal ambiguity is considered. As stated earlier this occurs when the windows of the channels are, in some sense, minimally spread.

Theorem 4.15. Given sufficient channel diversity and modes in the signals, e.g., the sufficient conditions of theorem 4.14, then nullity($\boldsymbol{\mathcal{I}}(\vartheta)$) = K^2 if and only if $L_j \in \{L_0, L_0+1\}$ for all $j = 1, \ldots, K$, for some integer L_0.

Proof. Under the conditions of theorem 4.14, then nullity($\boldsymbol{\mathcal{I}}(\vartheta)$) = NLB and from the proof of corollary 4.8, NLB = K^2 if and only if $(L_i - L_j + 1)_+ + (L_j - L_i + 1)_+ \equiv 2$ for each $i, j = 1, \ldots, K$. Either $(L_i - L_j + 1)_+ = (L_j - L_i + 1)_+ = 1$ or $(L_i - L_j + 1)_+ = 2$ and $(L_j - L_i + 1)_+ = 0$ or vice versa. □

The next corollary details the necessary and sufficient data size for the sources under the special case of equivalent channel orders.

Corollary 4.16. Given sufficient signal modes and channel diversity, e.g., as in theorem 4.14, if $L_k = L$ for each k and $M = K + 1$, then the CFIM attains the NLB if and only if

$$N \geq (K+2)KL + K.$$

4.1.4 Constraints for the convolutive mixture model

In the previous section, necessary and sufficient conditions on characteristics of the signal source and channel properties were detailed for the CFIM to have a null space with minimal dimensions equaling the nullity lower bound $\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+$. In this section, pathways to regularity in the CFIM (and hence the FIM) as well as several typical constraint sets are considered.

4.1.4.1 Pathways to regularity

Requiring constraints to attain regularity in the Fisher information is unnecessary to use this CCRB theory (theorem 3.17), but often without constraints the identifiable parameters are not in the desired framework to be of use to the practitioner. In the communications context, constraints on a model assumed a priori are simply a common method used to maintain a particular parametric structure in the model.

Guided by theorems 3.23, 3.24, 3.25, 3.27, and 3.29, then an objective of communications system design is to discover constraints for which identifiability of the parameter, as they are defined, is achieved, i.e., what properties can be imposed on the signal or channel to guarantee parametric identifiability. For the convolutive mixture model in (4.2), this involves an examination of the Fisher information in (4.10) with the CFIM in (4.12). In some sense, this has already been done. In section 4.1.3.3, conditions under which the CFIM has a null space spanned by the columns of the matrix $\boldsymbol{\mathcal{N}}$, defined in (4.14), are derived.

If $\boldsymbol{I}(\boldsymbol{\theta})$ is a Fisher information with Cholesky factorization $\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}^T(\boldsymbol{\theta})$ where $\boldsymbol{V}(\boldsymbol{\theta})$ is an orthogonal complement to $\boldsymbol{L}^T(\boldsymbol{\theta})$ [20, p.194]. Then theorem 3.23 implies for any constraint to achieve local identifiability, its Jacobian must satisfy $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{A}\boldsymbol{L}^T(\boldsymbol{\theta}) + \boldsymbol{B}\boldsymbol{V}^T(\boldsymbol{\theta})$ where $\boldsymbol{B}\boldsymbol{V}^T$ has full row rank. Similarly, theorem 3.24 implies the constraints must have an orthonormal complement $\boldsymbol{U}(\boldsymbol{\theta})$ satisfying $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{C} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{D}$ where $\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{C}$ if full column rank.

4.1.4.2 Norm channel + real-valued source constraint

The norm channel constraint

$$\left\|\boldsymbol{h}^{(1)}\right\|^2 = 1 \tag{4.16}$$

is a common scaling "trick" used in SIMO ($K = 1$ source) models to obtain channel estimates under second-order statistics assumptions. Combined with the (rotational) constraint of restricting the source elements to be real-valued, i.e.,

$$\mathrm{Im}(s^{(1)}(n)) = 0 \tag{4.17}$$

for $n = -L_1, -L_1+1, \ldots, N-1$, it is clear that the multiplicative ambiguity which is the basis of $\boldsymbol{\mathcal{N}}$ is eliminated. For the parameter vector $\boldsymbol{\theta} = \left[\mathrm{Re}(\boldsymbol{\vartheta}^T), \mathrm{Im}(\boldsymbol{\vartheta}^T)\right]^T$ with $\boldsymbol{\vartheta}$ defined in (4.11), the constraints are essentially separated as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{F}_{\mathrm{Re}(\boldsymbol{h})} & \boldsymbol{0} & \boldsymbol{F}_{\mathrm{Im}(\boldsymbol{h})} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}_{\mathrm{Re}(\boldsymbol{s})} & \boldsymbol{0} & \boldsymbol{F}_{\mathrm{Im}(\boldsymbol{s})} \end{bmatrix}$$

where $\boldsymbol{h} = \boldsymbol{h}^{(1)}$ and $\boldsymbol{s} = \boldsymbol{s}^{(1)}$. An orthonormal complement, or a basis for the null space, of $\left[\boldsymbol{F}_{\mathrm{Re}(\boldsymbol{s})}\ ,\ \boldsymbol{F}_{\mathrm{Im}(\boldsymbol{s})}\right]$ is simply $\begin{bmatrix} \boldsymbol{I}_{(N+L_1)\times(N+L_1)} \\ \boldsymbol{0} \end{bmatrix}$, but an analytic formula for the null space relating to the channel components is not so simple and requires numerical programming for arbitrary sizes L_1 and N.

Example 4.17. As a particular example of this constraint, consider the $M = 2$ SIMO convolutive channel model with $L = 3$ taps, where the channel is predefined by

$$\begin{bmatrix} \boldsymbol{h}_1^{(1)T} \\ \boldsymbol{h}_2^{(1)T} \end{bmatrix} = \begin{bmatrix} 0.3079 + j0.0698 & 0.1657 + j0.2304 & 0.0198 - j0.3823 & 0.0929 - j0.1853 \\ -0.1841 + j0.3294 & 0.4484 - j0.1689 & 0.0156 + j0.1526 & 0.4750 - j0.0952 \end{bmatrix},$$

under a white noise assumption (with known variance σ^2) and the constraints in (4.16) and (4.17). The 54 ($N = 50$) signal elements are BPSK (± 1) symbols randomly generated with equal probability (i.e., Bernoulli($\frac{1}{2}$)) and fixed for the simulation. The subspace method [52, 68] is used with a smoothing factor of 8. Additionally scoring with constraints (CSA) using (3.36) is applied on the subspace method's estimate $\hat{\boldsymbol{h}}^{(1)}$ for possible improvement. The results are shown in figure 4.2, where the mean-square error (MSE) per real channel coefficient is evaluated by $\frac{1}{16R}\sum_{r=1}^{R}\left\|\hat{\boldsymbol{h}}^{(1)} - \boldsymbol{h}^{(1)}\right\|^2$ over $R = 100$ runs or trials and the SNR is given by $10\log_{10}\frac{1}{\sigma^2}$. This example is also in [68] and shows how the subspace method, which had no prior theory-based performance metric, tracks with the CCRB. The additional overlaying estimation using the method of scoring demonstrates the local maximum likelihood properties in the subspace method.

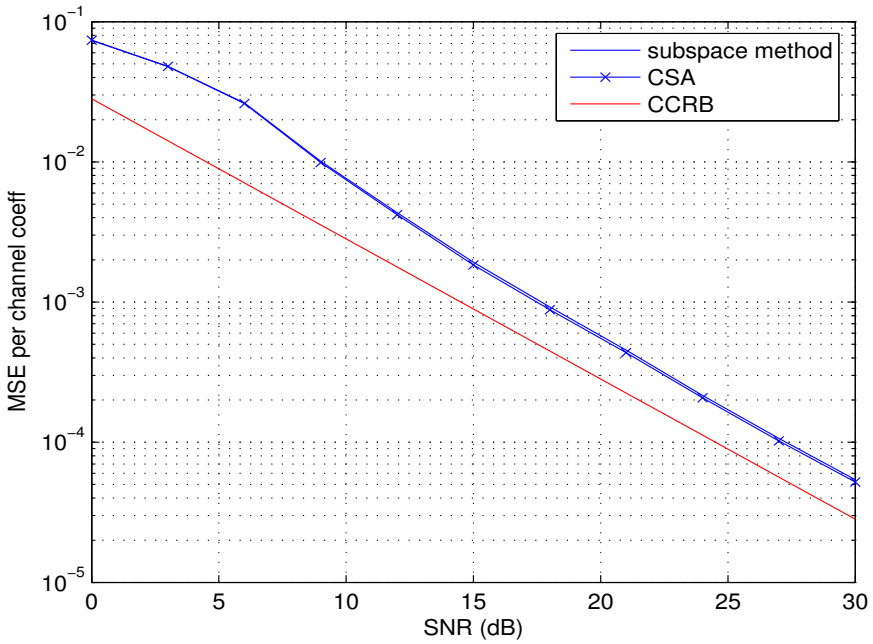


Figure 4.2: Norm-constrained channel estimation performance.

4.1.4.3 Semiblind constraints: $s^{(k)}(t) = p(t)$ for $t \in \mathbb{T}$

For multiple sources, designing communications with channel constraints, such as in the previous section, is less tenable since it is not always possible to establish guarantees on functions of the channel elements. The sources, however, are often entirely designable elements by restricting the class of signals which are to be passed through the channels. The class only needs to be defined by some functional constraint.

Perhaps, the simplest constraint is knowledge of a parameter element, i.e., $\theta_i = a$. This constraint produces a row vector $\boldsymbol{e}_i^T$ in the Jacobian $\boldsymbol{F}(\boldsymbol{\theta})$, where $\boldsymbol{e}_i$ is the unit vector with unity in the ith position and zero values in the other positions. Therefore, the corresponding orthonormal complement $\boldsymbol{U}(\boldsymbol{\theta})$ for this unit vector eliminates or nulls out the i row and column of the Fisher information $\boldsymbol{I}(\boldsymbol{\theta})$ while preserving the other elements. Any other nonredundant constraint in $\boldsymbol{f}$ does not change this.

Theorem 4.18. Assume the conditions of theorem 4.14, then nullity$(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) = \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+$. Let $\boldsymbol{\mathcal{I}}_*(\boldsymbol{\vartheta})$ be denoted by $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ with the rows $\{i_p : p = 1, \ldots, \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+\}$ and the corresponding columns removed. Then the nullity$(\boldsymbol{\mathcal{I}}_*(\boldsymbol{\vartheta})) = 0$ if and only if $\boldsymbol{\mathcal{N}}_*$ is full column rank, where $\boldsymbol{\mathcal{N}}_*$ is the matrix formed by taking the rows $\{i_p\}$ of $\boldsymbol{\mathcal{N}}$ in (4.14).

Proof. Let $\boldsymbol{F}$ be a matrix that when multiplied by the matrix $\boldsymbol{\mathcal{N}}$ selects the rows $\{i_p\}$, i.e., $\boldsymbol{F}$ is the matrix consisting of the row vectors $\{\boldsymbol{e}_i^T : i \in \{i_p\}\}$. Then $\boldsymbol{F}\boldsymbol{\mathcal{N}} = \boldsymbol{\mathcal{N}}_*$, a square full rank matrix. Therefore $\boldsymbol{F} = \boldsymbol{A}\boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{B}\boldsymbol{\mathcal{N}}^T$ for some

$\boldsymbol{A}$ and some $\boldsymbol{B}$ where $\boldsymbol{B}\boldsymbol{\mathcal{N}}^T$ is full row rank. As discussed in section 4.1.4.1, this identifies the parameters and hence nullity$(\boldsymbol{\mathcal{I}}_*(\boldsymbol{\vartheta})) = 0$ by theorem 3.24. □

This theorem can be used to specify either channel or source parameters to achieve identifiability and FIM regularity. For example, if only source signal parameters are specified, then it is necessary and sufficient to specify $\sum_{j=1}^{K}(L_i - L_j + 1)_+$ parameters of $\boldsymbol{s}^{(i)}$ for each $i = 1, \ldots, K$, under the conditions of theorem 4.18.

4.1.4.4 Unit Modulus constraint + Semiblind constraint

The unit or constant modulus constraint is a particularly powerful and reasonable assumption. All the source elements are assumed to have unit modulus, i.e.,

$$|s^{(k)}(n)|^2 = 1, \tag{4.18}$$

for every $n = -L_k, \ldots, N-1$ and for each $k = 1, \ldots, K$. This constraint is useful for modeling P-ary phase-shift keying (PSK) in communications models with unknown P, where the signals are assumed to be derived from a finite constellation of P equispaced points on the unit circle. While in practice, this assumption is often viewed as a single constraint, it is ultimately $\sum_{i=1}^{K}(N + L_i)$ constraints. Nevertheless, despite this, it is insufficient to identify the parameters by itself. For the parameter vector in (4.11), the constraint has a gradient

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{0} & 2\mathrm{Re}(\boldsymbol{S}_d^{(1)}) & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & 2\mathrm{Im}(\boldsymbol{S}_d^{(1)}) & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 2\mathrm{Re}(\boldsymbol{S}_d^{(2)}) & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & 2\mathrm{Im}(\boldsymbol{S}_d^{(2)}) & \cdots \\ & & & & \ddots & & & & & \vdots \end{bmatrix},$$

which has an orthonormal complement

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{M(L_1+1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ -\mathrm{Im}(\boldsymbol{S}_d^{(1)}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{M(L_2+1)} & \mathbf{0} & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathrm{Im}(\boldsymbol{S}_d^{(2)}) & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & & & & \ddots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{M(L_1+1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathrm{Re}(\boldsymbol{S}_d^{(1)}) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{M(L_2+1)} & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathrm{Re}(\boldsymbol{S}_d^{(2)}) & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & & & & \ddots \end{bmatrix}$$

satisfying (3.6) and where $\boldsymbol{S}_d^{(i)} = \mathrm{diag}(\boldsymbol{s}^{(i)})$. Using this complement with (4.10) and (4.12), we have that the i, j subblock of $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ has the structure $\boldsymbol{C}(i,j) =$

$$\begin{bmatrix} \mathrm{Re}[\boldsymbol{S}_d^{(i)H}\boldsymbol{H}_M^{(i)H}\boldsymbol{H}_M^{(j)}\boldsymbol{S}_d^{(j)}] & \mathrm{Im}[\boldsymbol{S}_d^{(i)H}\boldsymbol{H}_M^{(i)H}(\boldsymbol{I}_m \otimes \boldsymbol{S}^{(j)})] & \mathrm{Re}[\boldsymbol{S}_d^{(i)H}\boldsymbol{H}_M^{(i)H}(\boldsymbol{I}_m \otimes \boldsymbol{S}^{(j)})] \\ -\mathrm{Im}[(\boldsymbol{I}_m \otimes \boldsymbol{S}^{(i)H})\boldsymbol{H}_M^{(j)}\boldsymbol{S}_d^{(j)}] & \mathrm{Re}[\boldsymbol{I}_m \otimes \boldsymbol{S}^{(i)H}\boldsymbol{S}^{(j)}] & \mathrm{Im}[\boldsymbol{I}_m \otimes \boldsymbol{S}^{(i)H}\boldsymbol{S}^{(j)}] \\ \mathrm{Re}[\boldsymbol{H}_M^{(i)H}\boldsymbol{H}_M^{(j)}\boldsymbol{S}_d^{(j)}] & \mathrm{Im}[\boldsymbol{H}_M^{(i)H}(\boldsymbol{I}_m \otimes \boldsymbol{S}^{(j)})] & \mathrm{Re}[\boldsymbol{H}_M^{(i)H}(\boldsymbol{I}_m \otimes \boldsymbol{S}^{(j)})] \end{bmatrix}.$$

Lack of identifiability is noted since

$$\boldsymbol{C}(k,k) \cdot \begin{bmatrix} \mathbf{1}_{N+L_k} \\ \mathrm{Im}(\boldsymbol{h}^{(k)*}) \\ \mathrm{Re}(\boldsymbol{h}^{(k)*}) \end{bmatrix} = \mathbf{0}$$

for each $k = 1, \ldots, K$. Careful examination determines these are the only vectors in the null space of $\boldsymbol{U}^T(\boldsymbol{\theta}\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$, resulting in the following theorem.

Theorem 4.19. Assume $\mathrm{nullity}(\boldsymbol{I}(\boldsymbol{\theta})) = 2\sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+$. If all sources are assumed to be unit modulus and one complex-valued parameter for each source is assumed known, then $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ will be regular (and the model locally identifiable).

Intuitively, the unit modulus constraint eliminates the convolutive mixture and intra-source multiplicative amplitude ambiguity leaving only an intra-source multiplicative phase ambiguity.

Example 4.20. Consider a $K = 2$ source-$M = 2$ sensor linear instantaneous mixing model ($L_k = 0$ for all sources), with $N = 30$ samples per source, under a $R = 3$-ray multipath subchannels, i.e., the spatial signature of the kth source is expressed as a weighted sum of steering vectors, i.e., $\boldsymbol{h}_k = \sum_{r=1}^{R} \beta_{kr} \boldsymbol{a}(\psi_{kr})$ where β_{kr} and ψ_{kr} are the complex-valued amplitude and the real-valued AOA of the rth multipath of the kth source (see Figure 4.3). The AOAs and corresponding amplitudes are $\{\psi-1, \psi, \psi+4\}$

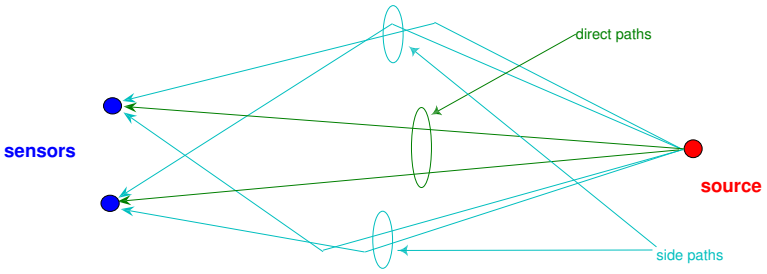


Figure 4.3: Example of multipath channel.

and $\{\sqrt{0.2}\angle(-\frac{\pi}{6}), \sqrt{0.5}, \sqrt{0.15}\angle(-\frac{pi}{5})\}$ for source $\boldsymbol{s}^{(1)}$ and $\{\psi+\Delta\psi-5, \psi+\Delta\psi, \psi+\Delta\psi+6\}$ and $\{\sqrt{0.15}\angle(-\frac{pi}{5}), \sqrt{0.6}, \sqrt{0.25}\angle(\frac{\pi}{3})\}$ for source $\boldsymbol{s}^{(2)}$. (See section 4.2 for a greater description on steering vectors.) The source elements come from an $8PSK$ alphabet with signal powers $\text{SNR}(\boldsymbol{s}^{(1)}) = 20\text{dB}$ and $\text{SNR}(\boldsymbol{s}^{(2)}) = 15\text{dB}$. The channel elements are normalized so that $\text{SNR}(\boldsymbol{s}^{(k)}) = \frac{\left\|\boldsymbol{h}^{(k)}\right\|^2}{M\sigma^2}$ with $\sigma^2 = 1$. The constraints assumed are unit modulus ($8PSK$) as well as knowledge of the first $T = 2$ symbols for each source, more than sufficient for identifiability and FIM regularity (theorem 4.19).

An initial estimate is obtained using the zero-forcing (bias reducing) variant of the algebraic constant modulus algorithm (ZF-ACMA) [71]. This algorithm is a useful tool for estimation of constant modulus source parameters in short data length

experiments (only $N \geq K^2$ or $2K$ required), but has no means of incorporating the training side information. The ZF-ACMA estimate is projected onto Θ_f to establish an initialization for the method of scoring with constraints. The step size rule chose $\alpha^{(k)} = 2^{-m}$ for the least positive integer m satisfying (3.37). The average MSE (per real parameter coefficient) at each iteration over 5000 trails is compared with the CCRB for each source in figure 4.4.

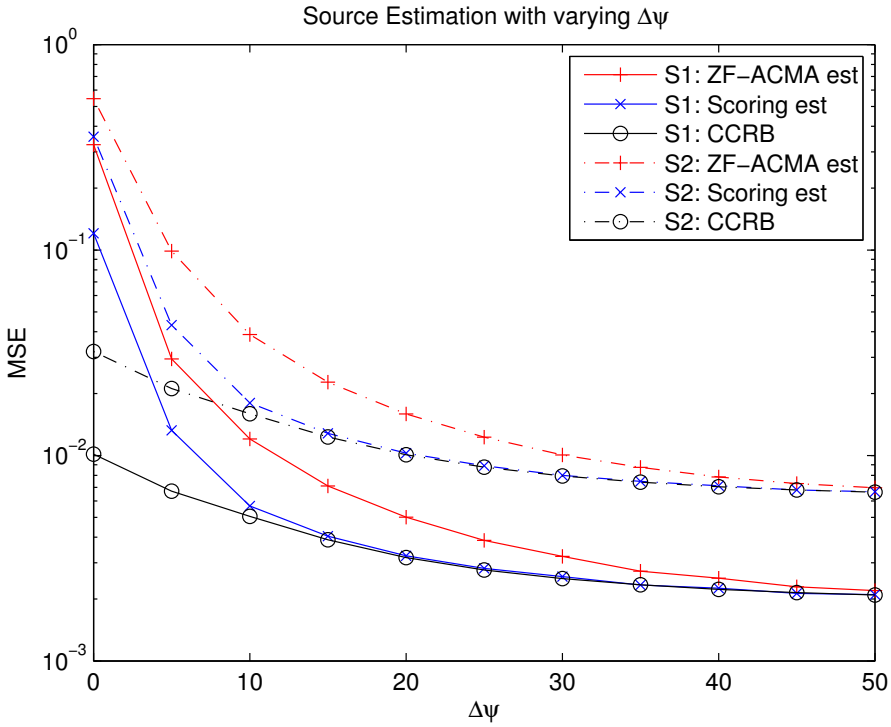


Figure 4.4: Source estimation with varying $\Delta\psi$.

The mean-square error improvement by utilizing the complete side information in scoring maintains efficiency with respect to the constrained CRB for moderately separated angle of arrivals compared with ZF-ACMA. As should be expected, the estimation performance degrades as the primary source angles overlap, but even in

the worst case scenario with $\Delta\psi = 0^\circ$ of separation, the approximate corresponding $8PSK$ bit-error-rate (BER) for ZF-ACMA and scoring with constraints is .2914 and .0254, meaning the estimation schemes result in bit decision errors roughly 29% and 3% of the time, respectively. The departure of the estimation performance from the CCRB as $\Delta\psi$ approaches 0° is possibly due to a loss of unbiasedness in the estimation.

4.2 Calibrated Array Model

The narrowband (calibrated) array model may be written as

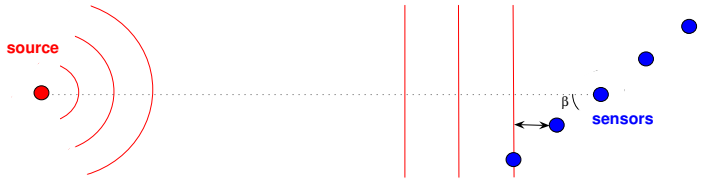
$$y_m(n) = \sum_{k=1}^{K} a_m(\omega_k)\gamma_k s^{(k)}(n) + w_m(n) \tag{4.19}$$

for $n = 1, \ldots, N$ and $m = 1, \ldots, M$, where $s^{(k)}(n)$ is the value of the kth input source at time index n, γ_k is the complex-valued channel gain for the kth input, $a_m(\omega_k)$ is the mth sensor response to the kth input source, ω_k is the angle-of-arrival (AOA) of the kth source, and $w_m(n)$ is the noise, modeled as zero-mean circular Gaussian with variance σ^2 iid in both time and space. In vector-matrix notation, the model for each time slot can be written as

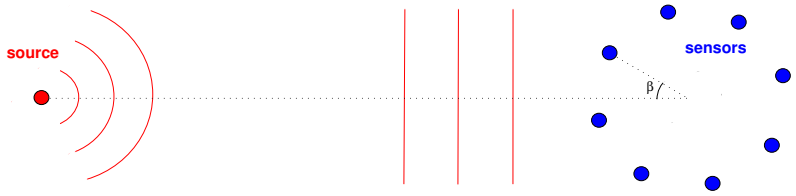
$$\boldsymbol{y}(n) = \sum_{k=1}^{K} \boldsymbol{a}(\omega_k)\gamma_k s^{(k)}(n) + \boldsymbol{w}(n) = \boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{\Gamma}\boldsymbol{s}(n) + \boldsymbol{w}(n) \tag{4.20}$$

where the input is given by $\boldsymbol{s}^T(n) = \left[s^{(1)}(n), \ldots, s^{(K)}(n)\right]$, the channel gain matrix is $\boldsymbol{\Gamma} = \text{diag}\left(\gamma_1, \ldots, \gamma_K\right)$, and the response matrix is given by

$$\boldsymbol{A}(\boldsymbol{\omega}) = \begin{bmatrix} \boldsymbol{a}(\omega_1) & \boldsymbol{a}(\omega_2) & \cdots & \boldsymbol{a}(\omega_K) \end{bmatrix} \tag{4.21}$$



(a) uniform linear array



(b) uniform circular array

Figure 4.5: Calibrated array model geometries for (a) uniform linear and (b) circular arrays.

with the array response vectors $\boldsymbol{a}(\omega_k)$ depending on the physical design of the array and the AOA. Calibration posits this known array geometry. The examples considered in this section consist of the uniform linear array (ULA) and uniform circular array (UCA), as shown in Figure 4.5. For example, for the ULA the response vectors typically have a Vandermonde vector structure with base $e^{-i\pi \sin(\omega_k)}$.

4.2.1 The Fisher information of the calibrated array model

4.2.1.1 Indirect derivation of the FIM

As was mentioned in section 4.1, the calibrated array model can be a special case of the convolutive mixture model using constraints. The method for transform-

ing the model is to keep the parameters for the instantaneous mixing model and to the parameter vector add extra parameters for the calibrated array and channel gain. When evaluating the FIM, the elements in the rows and columns corresponding to these extra parameters are zero. The constraint then represents the model reparameterization, e.g., for this case, we choose the constraints which define $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{\Gamma}$, element by element. The resulting CCRB submatrix corresponding to the extra parameters will be equivalent to the CRB of those parameters. This procedure is made clear in the following example.

Example 4.21. Assume $x \sim \mathcal{N}(ab, 1)$. The FIM for $\boldsymbol{\theta}^T = [a, b]$ is $\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} b^2 & ab \\ ab & a^2 \end{bmatrix}$, which is singular. Suppose we wish to reparameterize the model replacing ab with c. This is equivalent to the constraint $f(\boldsymbol{\theta}^\star) = ab - c$ for the expanded parameter vector $\boldsymbol{\theta}^{*T} = [a, b, c]$. For this orthonormal complement

$$\boldsymbol{U}(\boldsymbol{\theta}^*) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ b & a \end{bmatrix}$$

of the Jacobian of the constraints, then $\boldsymbol{U}^T(\boldsymbol{\theta}^*)\boldsymbol{I}(\boldsymbol{\theta}^*)\boldsymbol{U}(\boldsymbol{\theta}^*) = \boldsymbol{I}(\boldsymbol{\theta})$, which is still singular. Hence $\text{CCRB}(\boldsymbol{\theta}^*) = \boldsymbol{U}(\boldsymbol{\theta}^*)\boldsymbol{I}^\dagger(\boldsymbol{\theta})\boldsymbol{U}^T(\boldsymbol{\theta}^*)$ where the pseudoinverse is $\boldsymbol{I}^\dagger(\boldsymbol{\theta}) = \frac{1}{(a^2+b^2)^2}\boldsymbol{I}(\boldsymbol{\theta})$, or

$$\text{CCRB}(\boldsymbol{\theta}^*) = \begin{bmatrix} b^2 & ab & b^3 + a^2 b \\ ab & a^2 & ab^2 + a^3 \\ b^3 + a^2 b & ab^2 + a^3 & b^4 + 2a^2 b^2 + a^4 \end{bmatrix}.$$

The component corresponding to $\text{CCRB}(c) = \frac{1}{(a^2+b^2)^2}(b^4 + 2a^2 b^2 + a^4) = 1$, which agrees with the value of $\text{CRB}(c)$ for the model $x \sim \mathcal{N}(c, 1)$.

Note that the original model need not be identifiable, nor does the replacement model. Additional constraints under either model can also be included in the

constraint function. The example verifies that even for the more difficult case it is possible to find the CRB and CCRB for the calibrated array model indirectly from the instantaneous mixing model using constraints. However, in the interests of clarity, in the next section the FIM and CCRB will be derived for the calibrated array model directly.

4.2.1.2 Direct derivation of the FIM

The calibrated model in (4.19) has a likelihood given by

$$\begin{aligned} & p(\boldsymbol{y}(1),\ldots,\boldsymbol{y}(N);\boldsymbol{\theta}) \\ &= \frac{1}{(\pi\sigma^2)^{MN}}\exp\left\{-\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(\boldsymbol{y}(n)-\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{\Gamma}\boldsymbol{s}(n)\right)^H\left(\boldsymbol{y}(n)-\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{\Gamma}\boldsymbol{s}(n)\right)\right\}. \end{aligned} \tag{4.22}$$

For clarity, since there exists a mixture of complex and real parameters requiring estimation, then using the parameter vector[7]

$$\boldsymbol{\theta} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{s}(1)) \\ \mathrm{Im}(\boldsymbol{s}(1)) \\ \vdots \\ \mathrm{Re}(\boldsymbol{s}(N)) \\ \mathrm{Im}(\boldsymbol{s}(N)) \\ \boldsymbol{\gamma} \\ \boldsymbol{\omega} \end{bmatrix}, \tag{4.23}$$

with $\boldsymbol{\gamma}^T = \begin{bmatrix}\gamma_1,\ldots,\gamma_K\end{bmatrix}$ and $\boldsymbol{\omega}^T = \begin{bmatrix}\omega_1,\ldots,\omega_K\end{bmatrix}$, the Fisher information matrix is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{M}_1 & \boldsymbol{L}_1 \\ \boldsymbol{0} & \boldsymbol{M} & & \boldsymbol{0} & \boldsymbol{M}_2 & \boldsymbol{L}_2 \\ \vdots & & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{M} & \boldsymbol{M}_N & \boldsymbol{L}_N \\ \boldsymbol{M}_1^T & \boldsymbol{M}_2^T & \cdots & \boldsymbol{M}_N^T & \boldsymbol{K}_{\boldsymbol{\gamma}} & \boldsymbol{L} \\ \boldsymbol{L}_1^T & \boldsymbol{L}_2^T & \cdots & \boldsymbol{L}_N^T & \boldsymbol{L}^T & \boldsymbol{K}_{\boldsymbol{\omega}} \end{bmatrix} \tag{4.24}$$

[7]It is also possible to include the noise variance parameter σ^2 in the parameter vector. However, this parameter decouples from the other parameters resulting in an optimistic CRB of $\frac{\sigma^4}{MN}$ [16, 59], and so is uninteresting to the results herein.

where

$$\boldsymbol{M} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\mathcal{M}}) & -\mathrm{Im}(\boldsymbol{\mathcal{M}}) \\ \mathrm{Im}(\boldsymbol{\mathcal{M}}) & \mathrm{Re}(\boldsymbol{\mathcal{M}}) \end{bmatrix} \quad , \quad \boldsymbol{\mathcal{M}} = \frac{2}{\sigma^2}\boldsymbol{\Gamma}^H \boldsymbol{A}^H(\boldsymbol{\omega})\boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{\Gamma}, \tag{4.25}$$

$$\boldsymbol{M}_n = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\mathcal{M}}_n) & -\mathrm{Im}(\boldsymbol{\mathcal{M}}_n) \\ \mathrm{Im}(\boldsymbol{\mathcal{M}}_n) & \mathrm{Re}(\boldsymbol{\mathcal{M}}_n) \end{bmatrix} \quad , \quad \boldsymbol{\mathcal{M}}_n = \frac{2}{\sigma^2}\boldsymbol{\Phi}^H \boldsymbol{A}^H(\boldsymbol{\omega})\boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{S}(n), \tag{4.26}$$

$$\boldsymbol{L}_n = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\mathcal{L}}_n) \\ \mathrm{Im}(\boldsymbol{\mathcal{L}}_n) \end{bmatrix} \quad , \quad \boldsymbol{\mathcal{L}}_n = \frac{2}{\sigma^2}\boldsymbol{\Gamma}^H \boldsymbol{A}^H(\boldsymbol{\omega})\boldsymbol{D}(\boldsymbol{\omega})\boldsymbol{\Gamma}\boldsymbol{S}(n), \tag{4.27}$$

$$\boldsymbol{L} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\mathcal{L}}) \\ \mathrm{Im}(\boldsymbol{\mathcal{L}}) \end{bmatrix} \quad , \quad \boldsymbol{\mathcal{L}} = \frac{2}{\sigma^2}\sum_{n=1}^{N} \boldsymbol{S}^H(n)\boldsymbol{A}^H(\boldsymbol{\omega})\boldsymbol{D}(\boldsymbol{\omega})\boldsymbol{\Gamma}\boldsymbol{S}(n), \tag{4.28}$$

$$\boldsymbol{K}_{\boldsymbol{\gamma}} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & -\mathrm{Im}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) \\ \mathrm{Im}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & \mathrm{Re}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) \end{bmatrix} \quad , \quad \boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}} = \frac{2}{\sigma^2}\sum_{n=1}^{N} \boldsymbol{S}^H(n)\boldsymbol{A}^H(\boldsymbol{\omega})\boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{S}(n), \tag{4.29}$$

$$\text{and } \boldsymbol{K}_{\boldsymbol{\omega}} = \frac{2}{\sigma^2}\sum_{n=1}^{N} \mathrm{Re}\left(\boldsymbol{S}^H(n)\boldsymbol{\Gamma}^H \boldsymbol{D}^H(\boldsymbol{\omega})\boldsymbol{D}(\boldsymbol{\omega})\boldsymbol{\Gamma}\boldsymbol{S}(n)\right). \tag{4.30}$$

In the equations above, we redefine $\boldsymbol{S}(n) = \mathrm{diag}\left(\boldsymbol{s}^{(1)}(n), \ldots, \boldsymbol{s}^{(K)}(n)\right)$ and

$$\boldsymbol{D}(\boldsymbol{\omega}) = \left[\frac{\partial \boldsymbol{a}(\omega_1)}{\partial \omega_1} \quad \frac{\partial \boldsymbol{a}(\omega_2)}{\partial \omega_2} \quad \cdots \quad \frac{\partial \boldsymbol{a}(\omega_K)}{\partial \omega_K}\right].$$

4.2.1.3 Properties of the FIM

The model in (4.19) admits an ambiguity as $\gamma_k s^{(k)}(n)$ is indistinguishable from $\left(c_k\gamma_k\right)\left(\frac{s^{(k)}(n)}{c_k}\right)$ for any nonzero $c_k \in \mathbb{C}$. From Section 2.2, then it should be expected that the FIM in (4.24) is singular. This is indeed the case, e.g., note that

$$\begin{bmatrix} \boldsymbol{\mathcal{M}} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{\mathcal{M}}_1 \\ \boldsymbol{0} & \boldsymbol{\mathcal{M}} & & \boldsymbol{0} & \boldsymbol{\mathcal{M}}_2 \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\mathcal{M}} & \boldsymbol{\mathcal{M}}_N \\ \boldsymbol{\mathcal{M}}_1^T & \boldsymbol{\mathcal{M}}_2^T & \cdots & \boldsymbol{\mathcal{M}}_N^T & \boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}} \\ \boldsymbol{\mathcal{L}}_1^T & \boldsymbol{\mathcal{L}}_2^T & \cdots & \boldsymbol{\mathcal{L}}_N^T & \boldsymbol{\mathcal{L}}^T \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{S}(1) \\ \boldsymbol{S}(2) \\ \vdots \\ \boldsymbol{S}(N) \\ -\boldsymbol{\Gamma} \end{bmatrix} = \boldsymbol{0}.$$

And since the FIM consists of real and imaginary parts of this matrix, it too has a null space, namely, the columns of

$$\boldsymbol{N}(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{Re}(\boldsymbol{S}(1)) & -\mathrm{Im}(\boldsymbol{S}(1)) \\ \mathrm{Im}(\boldsymbol{S}(1)) & \mathrm{Re}(\boldsymbol{S}(1)) \\ \vdots & \vdots \\ \mathrm{Re}(\boldsymbol{S}(N)) & -\mathrm{Im}(\boldsymbol{S}(N)) \\ \mathrm{Im}(\boldsymbol{S}(N)) & \mathrm{Re}(\boldsymbol{S}(N)) \\ -\mathrm{Re}(\boldsymbol{\Gamma}) & \mathrm{Im}(\boldsymbol{\Gamma}) \\ -\mathrm{Im}(\boldsymbol{\Gamma}) & -\mathrm{Re}(\boldsymbol{\Gamma}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

are a basis for the null space (or at least a null subspace) of $\boldsymbol{I}(\boldsymbol{\theta})$. So the nullity, or dimension of the null space, of $\boldsymbol{I}(\boldsymbol{\theta})$ is at least $2K$. In fact, it is exactly $2K$ (provided $N \geq \frac{K}{M-K}$ for reasons similar to that given in theorem 4.12).

4.2.2 Constraints for the calibrated array model

4.2.2.1 Constraints on the complex-valued channel gain: $\boldsymbol{\Gamma} = \boldsymbol{I}_{K\times K}$

One approach to eliminating the ambiguity between the complex-valued gain and the source input is to incorporate this gain into the signal. Instead of remodeling the mean of (4.19) to be

$$\boldsymbol{\mu}(n, \boldsymbol{\theta}) = \sum_{k=1}^{K} \boldsymbol{a}(\omega_k) s^{(k)}(n) \tag{4.31}$$

by eliminating the unknown gain, it is equivalent to impose the constraints $\gamma_k = 1$ for $k = 1, \ldots, K$. The model in (4.31) is a model presented in a paper on "direction finding with narrow-band sensor arrays" by Stoica and Nehorai [67]. Hence, if the theory of Chapter 3 is to be trusted, then results found by imposing proper constraints that $\boldsymbol{\Gamma} = \boldsymbol{I}_{K\times K}$ (or $\boldsymbol{\gamma} = \mathbf{1}_K$) should be equivalent to the results of Stoica

and Nehorai. The K constraints of a complex-valued parameter $\boldsymbol{\gamma}$ are $2K$ constraints of the corresponding real-valued parameters, i.e., the vector $\boldsymbol{f}$ of constraints can be defined as

$$\begin{aligned} f_k(\boldsymbol{\theta}) &= \mathrm{Re}(\gamma_k) - 1 = 0 \\ f_{K+k}(\boldsymbol{\theta}) &= \mathrm{Im}(\gamma_k) = 0 \end{aligned}$$

for $k = 1, \ldots, K$. For the parameter vector in (4.23), the Jacobian of these constraints is given by

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{0}_{2K\times 2KN} & \boldsymbol{I}_{2K\times 2K} & \mathbf{0}_{2K\times K} \end{bmatrix}.$$

Since $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{N}(\boldsymbol{\theta}) = \boldsymbol{I}_{2K\times 2K}$, then by theorem 3.23, this constraint is sufficient to (locally) identify the parameters and by theorem 3.24 the matrix $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ will be regular for any matrix $\boldsymbol{U}(\boldsymbol{\theta})$ satisfying (3.6). An orthonormal basis for the null space of the Jacobian would be the columns of

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{I}_{2KN\times 2KN} & \mathbf{0}_{2KN\times K} \\ \mathbf{0}_{2K\times 2KN} & \mathbf{0}_{2K\times K} \\ \mathbf{0}_{K\times 2KN} & \boldsymbol{I}_{K\times K} \end{bmatrix},$$

and this generates a reduced "FIM"

$$\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{M} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{L}_1 \\ \mathbf{0} & \boldsymbol{M} & & \mathbf{0} & \boldsymbol{L}_2 \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{M} & \boldsymbol{L}_N \\ \boldsymbol{L}_1^T & \boldsymbol{L}_2^T & \cdots & \boldsymbol{L}_N^T & \boldsymbol{K}_{\boldsymbol{\omega}} \end{bmatrix},$$

which is equivalent to the Fisher information in [67, equation (E.9)]. Since the $\boldsymbol{\gamma}$ parameters are known and therefore it is unnecessary to understand the performance potential (or CCRB) of a known parameter, it is only of interest to have a bound

on the performance of estimators of the transformation

$$\boldsymbol{\alpha} = \boldsymbol{k}(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{Re}(\boldsymbol{s}(1)) \\ \mathrm{Im}(\boldsymbol{s}(1)) \\ \vdots \\ \mathrm{Re}(\boldsymbol{s}(N)) \\ \mathrm{Im}(\boldsymbol{s}(N)) \\ \boldsymbol{\omega} \end{bmatrix},$$

which has the Jacobian $\frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\alpha} = \boldsymbol{U}^T(\boldsymbol{\theta})$. Hence, $\mathrm{CCRB}(\boldsymbol{\alpha}) = \left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}$ is the same as the CRB found by Stoica and Nehorai. This equivalence serves as further validation of the CCRB approach.

4.2.2.2 Semiblind constraints: $\boldsymbol{s}(t) = \boldsymbol{p}(t)$ for $t \in \mathbb{T}$

An alternative approach to eliminating the ambiguity between the source and coefficient is the have prior knowledge of some of the source signals. These known elements are often referred to as *training* or *pilot symbols* in communications. Knowledge of any kth source element $s^{(k)}(t) = p^{(k)}(t)$ at any time sample t resolves the ambiguity between $\gamma_k s^{(k)}(n)$ for all n since γ_k is solvable in the observation corresponding to time sample t and can thus be used to solve the unknown source values when $n \notin \mathbb{T}$. This model can be written as

$$\boldsymbol{\mu}(n, \boldsymbol{\theta}) = \sum_{k=1}^{K} \boldsymbol{a}(\omega_k) \left(\gamma_k p^{(k)}(n)\delta_{n\in\mathbb{T}} + \gamma_k s^{(k)}(n)\delta_{n\notin\mathbb{T}}\right) \tag{4.32}$$

where $\delta_{\text{statement}} = 1$ when the statement is true and $= 0$ when the statement is false. This model is equivalent to the model designed by Kozick and Sadler [39, equation (9)] except with $\gamma_k s^{(k)}(n)$ being simply $s^{(k)}(n)$, a distinction that still allows for a match of results of the CCRB of a properly chosen transformation. The model is also equivalent to the model designed by Li and Compton [41, equation (2)] when

$\mathbb{T} = \{1, 2, \ldots, N\}$. Equivalence of the reparameterized CRBs in [39, 41] with the CCRB can be found in [59]. As in the previous example, the CCRB on the unknown parameters is the inverse of the (unconstrained) FIM after the elimination of its rows and columns corresponding to the known or specified parameters.

4.2.2.3 Finite alphabet constraint: $s^{(k)}(n) \in \mathbb{S}$

This constraint derives from the assumption that the source elements exists in a (finite) discrete set $\mathbb{S}$. In communications models, this corresponds to digital modulation designs such as pulse amplitude modulation (PAM), quadrature amplitude modulation (QAM), phase-shift keying (PSK), etc. As such, the model can also be the same as any of the previous models in (4.19), (4.31), or (4.32), but the defining characteristic is that the source samples only exist on a discrete set. There do not exist any CRB-type bounds for this model in the literature due to the problem of differentiability with respect to the parameters.[8] It is certainly possible to constrain any single real-valued parameter to a discrete set by creating a polynomial (or sine function) whose zeros match the set values. But what information can be gained from a constraint formulation of this discrete-alphabet model? This is perhaps best answered with the following example.

Example 4.22. Reconsider the model in example 2.2, $y \sim \mathcal{CN}(\vartheta, \sigma^2)$. Suppose the

[8] A Chapman-Robbins or Barankin-type bound, which does not require differentiability with respect to the parameters would be possible but even this approach does not seem to exist in the literature. Many communications engineers also discount the importance of a mean-square error bound for a digital signal and rely an alternative performance criteria, such as bit-error rate (BER).

parameter is required to satisfy the constraints

$$\begin{aligned} f_1(\vartheta) &= (\mathrm{Re}(\vartheta))^2 - \frac{1}{2} = 0 \\ f_2(\vartheta) &= (\mathrm{Im}(\vartheta))^2 - \frac{1}{2} = 0. \end{aligned}$$

This constraint restricts the real and imaginary part of ϑ to reside in the discrete set $\{\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\}$, i.e.

$$\vartheta \in \mathbb{S} = \left\{ \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2} \right\}.$$

(In the communications vernacular, this is quadrature PSK or 4-QAM.) The Jacobian for this constraint is

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} 2\mathrm{Re}(\vartheta) & 0 \\ 0 & 2\mathrm{Im}(\vartheta) \end{bmatrix},$$

which is full column rank for the possible set of values for ϑ. Hence, to satisfy (3.6), $\boldsymbol{U}(\boldsymbol{\theta}) = [\,]$ is a 2×0 null matrix.

This is analogous to the determinate case of example 3.14, therefore knowledge that a parameter exists in a discrete set is equivalent to complete knowledge of the parameter value in terms of mean-square error performance potential-the result being a Cramér-Rao bound of zero. This does not mean that the mean-square error will be zero (the decision of which set value the parameter actually is can be wrong given sufficient power in the noise), but it does mean that the mean-square error bound is trivial and not particularly helpful. This result is not surprising considering the degrees of freedom of the parameter that are restricted from such constraints. The restriction of a real-valued parameter to satisfying a single root equation eliminates its single degree of freedom.

4.2.2.4 Unit modulus constraints: $|\boldsymbol{s}(n)| = 1$ for all n

Since knowledge of parameters from a discrete (finite) alphabet results in a trivial bound, then to obtain a relevant and useful measure of performance potential a relaxation of the side information is necessary. One such example of this approach is using a constant or unit modulus constraint on the source elements as in section 4.1.4.4 for the convolutive mixture model. This approach remodels the mean as

$$\boldsymbol{\mu}(n, \boldsymbol{\theta}) = \sum_{k=1}^{K} \boldsymbol{a}(\omega_k)\gamma_k e^{j\phi^{(k)}(n)}. \tag{4.33}$$

This is the constraint considered in example 3.4 applied to the communications context. Therefore, imposing the constraints

$$f_{(n-1)K+k}(\boldsymbol{\theta}) = \left|s^{(k)}(n)\right|^2 - 1 = 0,$$

for $k = 1, \ldots, K$, $n = 1, \ldots, N$, is an alternative approach than rederiving the Fisher information for the model in (4.33). The Jacobian for this constraint has the form

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} 2\mathrm{Re}(\boldsymbol{S}(1)) & 2\mathrm{Im}(\boldsymbol{S}(1)) & & & \boldsymbol{0}_{K\times 3K} \\ & & \ddots & & \vdots \\ & & & 2\mathrm{Re}(\boldsymbol{S}(N)) & 2\mathrm{Im}(\boldsymbol{S}(N)) \quad \boldsymbol{0}_{K\times 3K} \end{bmatrix}, \tag{4.34}$$

which has a null space generated by the columns of the matrix

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} -\mathrm{Im}(\boldsymbol{S}(1)) & & & \\ \mathrm{Re}(\boldsymbol{S}(1)) & & & \\ & \ddots & & \\ & & -\mathrm{Im}(\boldsymbol{S}(N)) & \\ & & \mathrm{Re}(\boldsymbol{S}(N)) & \\ & & & \boldsymbol{I}_{3K\times 3K} \end{bmatrix}. \tag{4.35}$$

From this we can check the (local) identifiability of this model under the (unit) constant modulus constraint using theorem 3.24. The matrix $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is

$$\begin{bmatrix} \mathrm{Re}(\boldsymbol{S}^*(1)\boldsymbol{\mathcal{M}}\boldsymbol{S}(1)) & & & \mathrm{Im}(\boldsymbol{S}^*(1)\boldsymbol{\mathcal{M}}_1) & \mathrm{Re}(\boldsymbol{S}^*(1)\boldsymbol{\mathcal{M}}_1) & \mathrm{Im}(\boldsymbol{S}^*(1)\boldsymbol{\mathcal{L}}_1) \\ & \ddots & & \vdots & \vdots & \vdots \\ & & \mathrm{Re}(\boldsymbol{S}^*(N)\boldsymbol{\mathcal{M}}\boldsymbol{S}(N)) & \mathrm{Im}(\boldsymbol{S}^*(N)\boldsymbol{\mathcal{M}}_N) & \mathrm{Re}(\boldsymbol{S}^*(N)\boldsymbol{\mathcal{M}}_N) & \mathrm{Im}(\boldsymbol{S}^*(N)\boldsymbol{\mathcal{L}}_N) \\ -\mathrm{Im}(\boldsymbol{\mathcal{M}}_1^H\boldsymbol{S}(1)) & \cdots & -\mathrm{Im}(\boldsymbol{\mathcal{M}}_N^H\boldsymbol{S}(N)) & \mathrm{Re}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & -\mathrm{Im}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & \mathrm{Re}(\boldsymbol{\mathcal{L}}) \\ \mathrm{Re}(\boldsymbol{\mathcal{M}}_1^H\boldsymbol{S}(1)) & \cdots & \mathrm{Re}(\boldsymbol{\mathcal{M}}_N^H\boldsymbol{S}(N)) & \mathrm{Im}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & \mathrm{Re}(\boldsymbol{\mathcal{K}}_{\boldsymbol{\gamma}}) & \mathrm{Im}(\boldsymbol{\mathcal{L}}) \\ -\mathrm{Im}(\boldsymbol{\mathcal{L}}_1^H\boldsymbol{S}(1)) & \cdots & -\mathrm{Im}(\boldsymbol{\mathcal{L}}_N^H\boldsymbol{S}(N)) & \mathrm{Re}(\boldsymbol{\mathcal{L}}^H) & -\mathrm{Im}(\boldsymbol{\mathcal{L}}^H) & \boldsymbol{\mathcal{K}}_{\boldsymbol{\omega}}. \end{bmatrix}$$

Since

$$\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \cdot \begin{bmatrix} \boldsymbol{I}_{K\times K} \\ \vdots \\ \boldsymbol{I}_{K\times K} \\ -\mathrm{Im}(\boldsymbol{\Gamma}) \\ \mathrm{Re}(\boldsymbol{\Gamma}) \\ \boldsymbol{0}_{K\times K} \end{bmatrix} = \boldsymbol{0},$$

then $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is singular and by theorem 3.24 the (unit) constant modulus constraints are not sufficient for identifiability. This was also the case for the convolutive mixture model in section 4.1.4.4. Furthermore, reviewing the model in (4.33), this result has more reason to be expected. The original identifiability issue in the calibrated model (4.19) is the multiplicative ambiguity between the sources and the channel gain. While the constant modulus constraint resolves any amplitude ambiguity of the channel gain, i.e., $|\gamma_k| = \left|\gamma_k s^{(k)}(n)\right|$ for any n, there still exist a phase rotation ambiguity. It is clear from both the model and from theorem 3.23, that for each source k knowledge of an element for either the real or imaginary part of either the channel gain or a source sample will be sufficient for identifiability and a regular $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$. Of course, given this constant unit modulus constraint, knowledge of the real (imaginary) part of any source sample is equivalent to restricting the imaginary (real) part of the source sample to a finite discrete alphabet, which as discussed in section 4.2.2.3 is the same as knowledge of the parameter in regards to the CCRB performance potential but not in regards to the estimation. Hence for estimation performance it is a necessity to constrain the real and/or imaginary part of the source sample.

4.2.2.5 Unit modulus constraint; real-valued channel gain: $\mathrm{Im}(\gamma_k) = 0$ for all k

This model is that given in (4.33) except with γ_k being real-valued. As seen in sections 4.2.2.2 and 4.2.2.3, the constraint is knowledge of the imaginary parts of $\boldsymbol{\gamma}$ and the effect of the corresponding $\boldsymbol{U}(\boldsymbol{\theta})$ matrix is the elimination of the columns and rows corresponding to the $\mathrm{Im}(\gamma_k)$ parameters. This model is equivalent to that of Leshem and van der Veen [40]. Verification that the CCRB is equivalent can be found in [59]. The bound is essentially the inverse of the *reduced-parameter-space Fisher information* $\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ from section 4.2.2.4 after the elimination of the rows and columns corresponding to the channel gain parameters.

4.2.2.6 Semi-blind and unit modulus constraint

This model merges the models in (4.33) and (4.32). Without loss of generality, assume the source elements are known for the first T time slots for each source. Then the constraint Jacobian is $\boldsymbol{F}(\boldsymbol{\theta}) =$

$$\begin{bmatrix} \boldsymbol{I}_{2TK\times 2TK} & & & & & \boldsymbol{0}_{2TK\times 3K} \\ & 2\mathrm{Re}(\boldsymbol{S}(T+1)) & 2\mathrm{Im}(\boldsymbol{S}(T+1)) & & & \boldsymbol{0}_{K\times 3K} \\ & & & \ddots & & \vdots \\ & & & 2\mathrm{Re}(\boldsymbol{S}(N)) & 2\mathrm{Im}(\boldsymbol{S}(N)) & \boldsymbol{0}_{K\times 3K} \end{bmatrix}, \tag{4.36}$$

which has an orthonormal complement

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{0}_{2TK\times K} & & & \\ -\mathrm{Im}(\boldsymbol{S}(1)) & & & \\ \mathrm{Re}(\boldsymbol{S}(1)) & & & \\ & \ddots & & \\ & & -\mathrm{Im}(\boldsymbol{S}(N)) & \\ & & \mathrm{Re}(\boldsymbol{S}(N)) & \\ & & & \boldsymbol{I}_{3K\times 3K} \end{bmatrix}. \tag{4.37}$$

The reduced FIM is then

$$\boldsymbol{U}^{(}\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{G}(T+1) & & & & \boldsymbol{C}(T+1) & \boldsymbol{B}(T+1) \\ & \boldsymbol{G}(T+2) & & & \boldsymbol{C}(T+2) & \boldsymbol{B}(T+2) \\ & & \ddots & \vdots & \vdots & \\ & & & \boldsymbol{G}(N) & \boldsymbol{C}(N) & \boldsymbol{B}(N) \\ \boldsymbol{C}^T(T+1) & \boldsymbol{C}^T(T+2) & \cdots & \boldsymbol{C}^T(N) & \boldsymbol{K}_{\boldsymbol{\gamma}} & \boldsymbol{L} \\ \boldsymbol{B}^T(T+1) & \boldsymbol{B}^T(T+2) & \cdots & \boldsymbol{B}^T(N) & \boldsymbol{L}^T & \boldsymbol{K}_{\boldsymbol{\omega}} \end{bmatrix} \tag{4.38}$$

where $\boldsymbol{G}(t) = \mathrm{Re}[\boldsymbol{S}^*(t)\boldsymbol{\mathcal{M}}\boldsymbol{S}(t)]$ and $\boldsymbol{\mathcal{M}}$ is defined as in (4.25), where $\boldsymbol{C}(t) = \begin{bmatrix}\mathrm{Im}[\boldsymbol{S}^*(t)\boldsymbol{\mathcal{M}}_t] & \mathrm{Re}[\boldsymbol{S}^*(t)\boldsymbol{\mathcal{M}}_t]\end{bmatrix}$ and $\boldsymbol{\mathcal{M}}_n$ is defined in (4.26), $\boldsymbol{B}(t) = \mathrm{Im}[\boldsymbol{S}^*(t)\boldsymbol{\mathcal{L}}_t]$ and $\boldsymbol{\mathcal{L}}_t$ is defined in (4.27). To analytically invert this matrix, the Schur complement,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{K}_{\boldsymbol{\gamma}} - \sum_{t=T+1}^{N} \boldsymbol{C}^T(t)\boldsymbol{G}^{-1}(t)\boldsymbol{C}(t) & \boldsymbol{L} - \sum_{t=T+1}^{N} \boldsymbol{C}^T(t)\boldsymbol{G}^{-1}(t)\boldsymbol{B}(t) \\ \boldsymbol{L}^T - \sum_{t=T+1}^{N} \boldsymbol{B}^T(t)\boldsymbol{G}^{-1}(t)\boldsymbol{C}(t) & \boldsymbol{K}_{\boldsymbol{\omega}} - \sum_{t=T+1}^{N} \boldsymbol{B}^T(t)\boldsymbol{G}^{-1}(t)\boldsymbol{B}(t) \end{bmatrix}, \tag{4.39}$$

is useful. If $\boldsymbol{\Phi}$ is partitioned into corresponding subblocks $\begin{bmatrix} \boldsymbol{\Phi}_{\boldsymbol{K}_\gamma} & \boldsymbol{\Phi}_{\boldsymbol{L}} \\ \boldsymbol{\Phi}_{\boldsymbol{L}^T} & \boldsymbol{\Phi}_{\boldsymbol{K}_\omega} \end{bmatrix}$, then the CCRB subblocks for the unknown elements are given by

$$\begin{aligned} \mathrm{CCRB}(\boldsymbol{\omega}) &= \left[\boldsymbol{\Phi}_{\boldsymbol{K}_\omega} - \boldsymbol{\Phi}_{\boldsymbol{L}^T}\boldsymbol{\Phi}_{\boldsymbol{K}_\gamma}^{-1}\boldsymbol{\Phi}_{\boldsymbol{L}}\right]^{-1} \\ \mathrm{CCRB}(\boldsymbol{\gamma}) &= \left[\boldsymbol{\Phi}_{\boldsymbol{K}_\gamma} - \boldsymbol{\Phi}_{\boldsymbol{L}}\boldsymbol{\Phi}_{\boldsymbol{K}_\omega}^{-1}\boldsymbol{\Phi}_{\boldsymbol{L}^T}\right]^{-1} \\ \mathrm{CCRB}(\begin{bmatrix}\mathrm{Re}(\boldsymbol{s}(t)) \\ \mathrm{Im}(\boldsymbol{s}(t))\end{bmatrix}) &= \begin{bmatrix}-\mathrm{Im}(\boldsymbol{S}(t)) \\ \mathrm{Re}(\boldsymbol{S}(t))\end{bmatrix} \boldsymbol{X}(t) \begin{bmatrix}-\mathrm{Im}(\boldsymbol{S}(t)) & \mathrm{Re}(\boldsymbol{S}(t))\end{bmatrix} \end{aligned}$$

where $\boldsymbol{X}(t) = \boldsymbol{G}^{-1}(t) + \boldsymbol{G}^{-1}(t)\begin{bmatrix}\boldsymbol{C}(t) & \boldsymbol{B}(t)\end{bmatrix}\boldsymbol{\Phi}^{-1}\begin{bmatrix}\boldsymbol{C}^T(t) \\ \boldsymbol{B}^T(t)\end{bmatrix}\boldsymbol{G}^{-1}(t)$.

Example 4.23. A distinct advantage of the CCRB, as implemented in this treatise, is the ability to compare the performance potential of a number of different models seamlessly. Suppose we consider $M = 5$ omni-directional sensors with a beamwidth

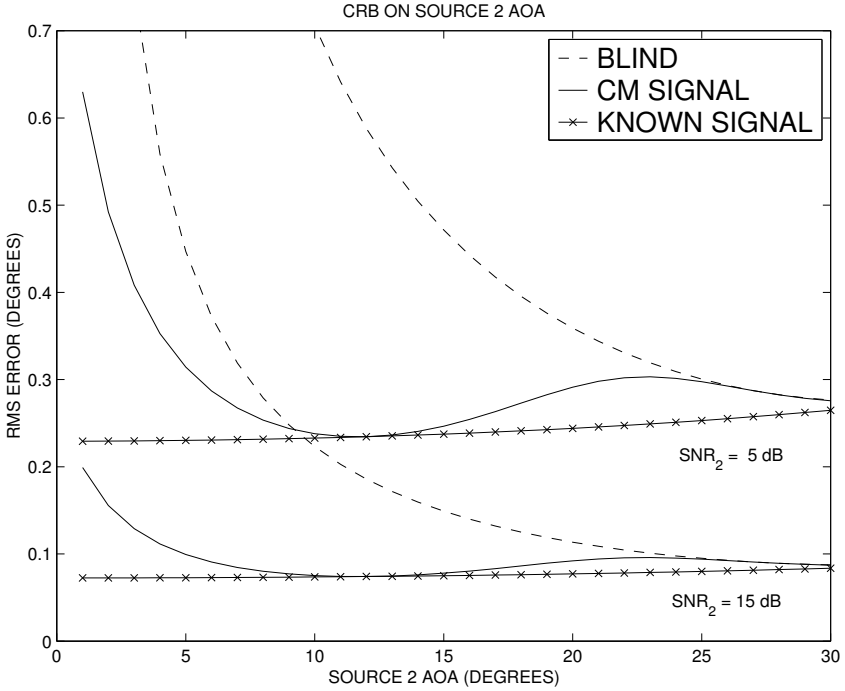


Figure 4.6: CCRBs on AOA for blind, constant modulus, and known signal models.

of $\approx 23.6^\circ$ in a uniform linear array receiving $K = 2$ source signals over $N = 100$ time samples. Then, it is possible to compare various communications design scenarios, e.g.,

(a) the "blind" case[9]: $s_k(1)$ known for $k = 1, 2$,

(b) the unit modulus case: $|s_k(t)|^2 = 1$ for $k = 1, 2$ and $t = 1, \ldots, 100$,

(c) the semiblind case: $s_k(t)$ known for $k = 1, 2$ and $t = 1, \ldots, T = 20$,

(d) the unit modulus and semiblind case: $s_k(t)$ known for $k = 1, 2$ and $t = 1, \ldots, T = 20$, and $|s_k(t)|^2 = 1$ for $k = 1, 2$ and $t = T + 1, \ldots, 100$, and

(e) the known signal case: $s_k(t)$ known for $k = 1, 2$ and $t = 1, \ldots, 100$.

[9]This is not a truly blind scenario, but is often referred to as blind in the literature.

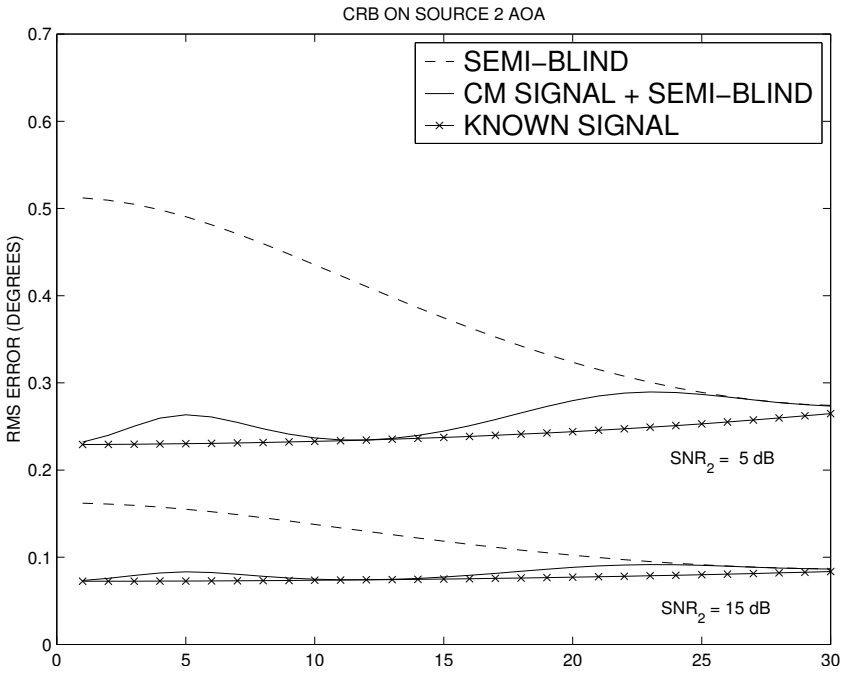


Figure 4.7: CCRBs on AOA for semiblind, constant modulus + semiblind, and known signal models.

Figure 4.6 displays a comparison between blind, unit modulus constraints, and known signal model designs of the CCRBs on angle-of-arrival (AOA) estimation for the second source signal over varying directions and different signal-to-noise ratios (SNRs) with the first source signal arriving at 0°. Figure 4.7 displays a comparison of CCRBs on AOA estimation between semiblind constraints, unit modulus with semiblind constraints, and known signal model designs. And finally, figure 4.8 displays CCRBs on signal phase estimation between blind, unit modulus constraints, semiblind constraints, and a mixture of unit modulus and semiblind constraints. The known signal model is the best case scenario for AOA estimation potential and is a useful guideline for more desirable scenarios where information (data or unknown parameters) is included in the transmission. Figures 4.6 and 4.7 demonstrate the characteristic loss of performance when the sources' AOAs differ by roughly the beamwidth. The value of the unit modulus constraint is evident when the sources' AOAs are closely spaced as the CCRB performance potential approximates that of the known signal model. In figures 4.7 and 4.8, the estimation potential actually improves for closely space sources with the semiblind and unit modulus constraint mixture.

4.3 Discussion

This chapter includes extensions of only a brief sampling of my prior research [59, 39, 48, 51, 39, 48, 59, 49] as it relates to the practical application of the CCRB. In this chapter, the convolutive mixture model and the calibrated array model were

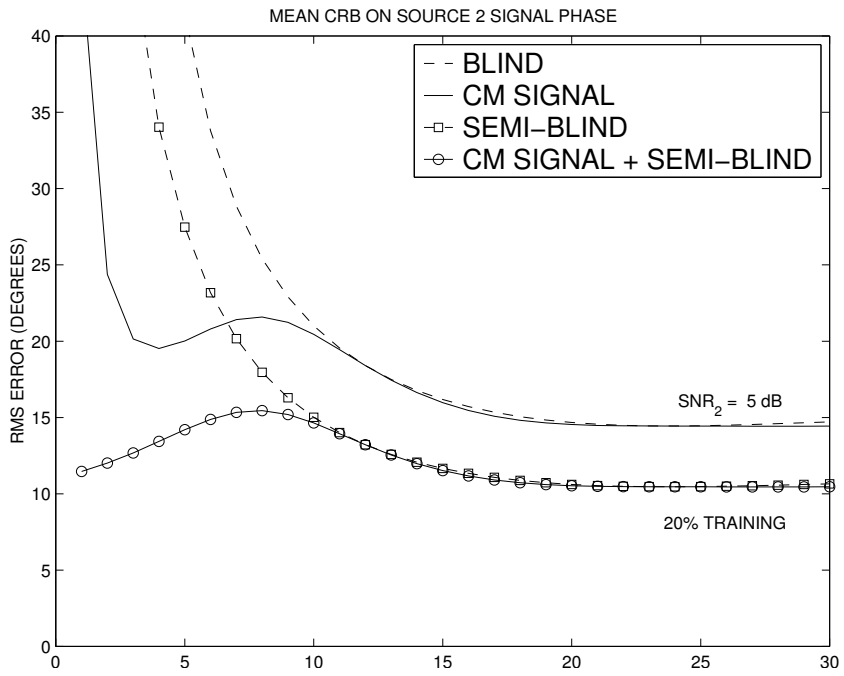


Figure 4.8: CCRBs on signal phase for blind, constant modulus, semiblind, and constant modulus + semiblind models.

treated as base models for which the Fisher information was derived a single time and then a series of variations on the models in the form of differentiable parametric constraints were considered. This approach presents a simple procedure to compare and contrast a large class of constraints, essentially different models, in an efficient manner to determine the value of particular formulation in terms of performance potential as measured in the CCRB.

Appendix A

Appendices

A.1 A proof of the CCRB using the Chapman-Robbins version of the Barankin bound

Gorman and Hero developed a CCRB using the multiparameter version of the Hammersley-Chapman-Robbins bound (HCRB) [16, 26]. However, the result produced a variant form of the CCRB

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^{T}(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^{T}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta}),$$

which requires a nonsingular FIM. What follows is a shorter variation of their approach that does not assume a nonsingular FIM, starting with a brief description of the HCRB.

Rather than relying on the Fisher score, which is the derivative of the log-likelihood, i.e.,

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})}\frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

the regularity conditions requiring a differentiable likelihood can be relaxed by considering finite differences, i.e., for each $i = 1, \ldots, m$,

$$\frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})}\frac{p(\boldsymbol{x};\boldsymbol{\theta} + \epsilon_i \boldsymbol{e}_i) - p(\boldsymbol{x};\boldsymbol{\theta})}{\epsilon_i}$$

where the $\boldsymbol{e}_i$ are canonical unit vectors. If the likelihood is differentiable then the limit as each $\epsilon_i \to 0$ is the Fisher score. Of course, the finite differences need not

be with respect to the canonical axis and the number of finite differences need not be the same as the dimension of the parameters. This is the generalization of the CRB known as the Hammersley-Chapman-Robbins (HCR) version of the Barankin bound.

Theorem A.1. If $\boldsymbol{t}(\boldsymbol{x})$ be an unbiased estimate of $\boldsymbol{h}(\boldsymbol{\theta}) \in \boldsymbol{h}(\mathbb{R}^m) \subset \mathbb{R}^t$, and $\boldsymbol{\psi} = \left[\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(p)}\right]$ is a matrix whose columns are test points in $\mathbb{R}^m$, all distinct from each other as well as from $\boldsymbol{\theta}$, then the variance of $\boldsymbol{t}(\boldsymbol{x})$ is bounded below by the inequality

$$\mathrm{Var}(\boldsymbol{t}(\boldsymbol{x})) \geq \sup_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(p)},p} \boldsymbol{\Delta}(\boldsymbol{\theta},\boldsymbol{\psi})\boldsymbol{\Upsilon}^{-1}(\boldsymbol{\theta},\boldsymbol{\psi})\boldsymbol{\Delta}^T(\boldsymbol{\theta},\boldsymbol{\psi})$$

where $\boldsymbol{\Delta}(\boldsymbol{\theta},\boldsymbol{\psi})$ is called a *translation matrix* defined by

$$\boldsymbol{\Delta}(\boldsymbol{\theta},\boldsymbol{\psi}) = \left[\boldsymbol{h}(\boldsymbol{\psi}^{(1)}) - \boldsymbol{h}(\boldsymbol{\theta}), \quad \cdots, \quad \boldsymbol{h}(\boldsymbol{\psi}^{(p)}) - \boldsymbol{h}(\boldsymbol{\theta})\right]$$

and $\boldsymbol{\Upsilon}(\boldsymbol{\theta},\boldsymbol{\psi})$ is called an *HCR information matrix* defined by

$$\boldsymbol{\Upsilon}_{ij}(\boldsymbol{\theta},\boldsymbol{\psi}) = E_{\boldsymbol{\theta}}\frac{p(\boldsymbol{x};\boldsymbol{\psi}^{(i)}) - p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})}\frac{p(\boldsymbol{x};\boldsymbol{\psi}^{(j)}) - p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})}.$$

This result encompasses the CRB result when the test points satisfy $\boldsymbol{\psi}^{(i)} = \boldsymbol{\theta} + \epsilon_i \boldsymbol{e}_i$ and the $\epsilon_i \to 0$ for $i = 1, \ldots, m = p$, i.e., as a properly chosen set of test points approach the parameter. If the set of vectors $\boldsymbol{\psi}^{(1)} - \boldsymbol{\theta}, \ldots, \boldsymbol{\psi}^{(m)} - \boldsymbol{\theta}$ span an m-dimensional space, then the limit as the finite differences approach derivatives will still obtain the CRB.

Since $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$ it makes sense to restrict the test points to also satisfy the constraints, $\boldsymbol{f}(\boldsymbol{\psi}^{(i)}) = \boldsymbol{0}$, and examine the limit of the HCRB as the finite differences

approach derivatives. The Taylor series approximation of $\boldsymbol{f}(\boldsymbol{\psi}^{(i)})$ about $\boldsymbol{\theta}$ is

$$\boldsymbol{f}(\boldsymbol{\psi}^{(i)}) = \boldsymbol{f}(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta})\left(\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}\right) + o(||\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}||).$$

Since $\boldsymbol{f}(\boldsymbol{\psi}^{(i)}) = \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, then $\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}$ almost entirely resides in null($\boldsymbol{F}(\boldsymbol{\theta})$). So without loss of generality we can allow

$$\boldsymbol{\psi}^{(i)} = \boldsymbol{\theta} + \delta_i \boldsymbol{u}_i(\boldsymbol{\theta}) + o(||\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}||)$$

where $\boldsymbol{u}_i(\boldsymbol{\theta})$ is the ith column of the matrix $\boldsymbol{U}(\boldsymbol{\theta})$ satisfying (3.6). Then

$$\begin{aligned}\frac{p(\boldsymbol{x};\boldsymbol{\psi}^{(i)}) - p(\boldsymbol{x};\boldsymbol{\theta})}{\delta_i p(\boldsymbol{x};\boldsymbol{\theta})} &= \frac{p(\boldsymbol{x};\boldsymbol{\theta} + \delta_i \boldsymbol{u}_i(\boldsymbol{\theta}) + o(||\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}||)) - p(\boldsymbol{x};\boldsymbol{\theta})}{\delta_i p(\boldsymbol{x};\boldsymbol{\theta})} \\ &\rightarrow \boldsymbol{u}_i^T(\boldsymbol{\theta})\frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})} = \boldsymbol{u}_i^T(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})\end{aligned}$$

and

$$\frac{\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}}{\delta_i} = \boldsymbol{u}_i(\boldsymbol{\theta}) + \frac{1}{\delta_i} o(||\boldsymbol{\psi}^{(i)} - \boldsymbol{\theta}||) \rightarrow \boldsymbol{u}_i(\boldsymbol{\theta})$$

as $\delta_i \rightarrow 0$. This is true for any i, so if the test points are chosen such that each column of $\boldsymbol{U}(\boldsymbol{\theta})$ is used, this gives us the CCRB

$$\boldsymbol{U}(\boldsymbol{\theta})\left(\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta})$$

as the limit of a constrained HCRB when the finite differences become derivatives.

A.2 A proof of the CCRB using the method of implicit differentiation

Suppose $\boldsymbol{\theta}$ is restricted to the zeros of $\boldsymbol{f} : \mathbb{R}^m \rightarrow \mathbb{R}^k$ with Jacobian $\boldsymbol{F}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ having rank k whenever $\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$. The method of implicit differentiation assumes that parameters that would be eliminated under a reparameterization can

be written in terms of the remaining parameters, since the conditions satisfy the implicit function theorem. The actual function generally remains unknown, but it's derivative is calculated by first taking partial derivatives of constraint function and using linear algebra to solve for the partial of the eliminated parameters in terms of the remaining parameters. This approach was also used by Marzetta [47, proof of theorem 1] to prove the regularity conditions given in (3.20).

The parameter vector may be separated as $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{bmatrix}$ and the constraint $\boldsymbol{f}$ may be rewritten as $\boldsymbol{f}^\star : \mathbb{R}^{m-k} \times \mathbb{R}^k \to \mathbb{R}^k$ via the mapping $\boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{f}(\begin{bmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{bmatrix})$. Then the Jacobian of $\boldsymbol{f}$ can be represented as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{f}^\star_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \boldsymbol{f}^\star_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\end{bmatrix}$$

where $\boldsymbol{f}^\star_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{\partial}{\partial \boldsymbol{\theta}_i^T}\boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for each $i = 1, 2$.

Without loss of generality, assume $\boldsymbol{\theta}_2 \in \mathbb{R}^k$ is a function of $\boldsymbol{\theta}_1 \in \mathbb{R}^{m-k}$, i.e., $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)$ is an implicit function. Therefore, $\boldsymbol{f}^\star$ is implicitly only a parameter of $\boldsymbol{\theta}_1$ and

$$\frac{\partial}{\partial \boldsymbol{\theta}_1^T}\boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)) = \boldsymbol{f}^\star_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)) + \boldsymbol{f}^\star_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1))\frac{\partial \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1^T}.$$

If this derivative is only taken where $\boldsymbol{f}^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{0}$, then in matrix form,

$$\begin{bmatrix}\boldsymbol{f}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \boldsymbol{f}_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_{m-k,m-k} \\ \frac{\partial \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}\end{bmatrix} = \boldsymbol{0}.$$

The first matrix above is $\boldsymbol{F}(\boldsymbol{\theta})$; the second matrix above consists of $m - k$ linearly independent columns which exist in the null space of the row vectors of $\boldsymbol{F}(\boldsymbol{\theta})$, i.e., the second matrix is merely some transformation of some matrix $\boldsymbol{U}(\boldsymbol{\theta})$ defined as in (3.6).

A.3 Alternative proof of asymptotic normality

Crowder [18] proved the following theorem.

Theorem A.2 (Crowder)**.** If

1. there is a consistent solution $(\dot{\boldsymbol{\theta}}_n, \dot{\boldsymbol{\lambda}}_n)$ of the likelihood equations,
2. $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}_0) \overset{d}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\theta}_0))$,
3. $\boldsymbol{D}^{-1}(\dot{\boldsymbol{\theta}}_n)\left(\boldsymbol{I}(\boldsymbol{\theta}_0) + \frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}_0,\dot{\boldsymbol{\theta}}_n)\right) \xrightarrow{p} \boldsymbol{0}$,
4. $\boldsymbol{Q}(\dot{\boldsymbol{\theta}}_n) - \boldsymbol{Q}(\boldsymbol{\theta}_0) \xrightarrow{p} \boldsymbol{0}$, and
5. $\det \boldsymbol{Q}(\dot{\boldsymbol{\theta}}_n) \leq K < \infty$
6. $\det \boldsymbol{D}^{-1}(\dot{\boldsymbol{\theta}}_n)\boldsymbol{F}^T(\dot{\boldsymbol{\theta}}_n)\left(\boldsymbol{F}(\boldsymbol{\theta},\dot{\boldsymbol{\theta}}_n)\boldsymbol{D}^{-1}(\dot{\boldsymbol{\theta}}_n)\boldsymbol{F}^T(\dot{\boldsymbol{\theta}}_n)\right)^{-1} \leq K < \infty$

where $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}_0,\boldsymbol{\theta})$ is a matrix in the form of the Fisher score $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})$ but with each row evaluated at possibly different points on the line between $\boldsymbol{\theta}_0$ and $\dot{\boldsymbol{\theta}}_n$, and similarly for $\boldsymbol{F}(\boldsymbol{\theta},\dot{\boldsymbol{\theta}}_n)$, and $\boldsymbol{Q}(\boldsymbol{\theta}) = \boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}$. Then

$$\sqrt{n}\left(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{D}^{-1}(\boldsymbol{\theta}) - \boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\left(\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\boldsymbol{F}^T(\boldsymbol{\theta})\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{D}^{-1}(\boldsymbol{\theta})\right)$$

where $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{K}\boldsymbol{F}(\boldsymbol{\theta})$ for an arbitrary positive semi-definite matrix $\boldsymbol{K}$.

Crowder's asymptotic normality theorem shows that variance of the CMLE satisfies

$$\mathrm{Var}(\sqrt{n}\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} \boldsymbol{D}^{-1}(\boldsymbol{\theta}_0) - \boldsymbol{D}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{F}^T(\boldsymbol{\theta}_0)\left(\boldsymbol{F}(\boldsymbol{\theta}_0)\boldsymbol{D}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{F}^T(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta}_0)\boldsymbol{D}^{-1}(\boldsymbol{\theta}_0)$$

as $n \to \infty$, where $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + \boldsymbol{F}^T(\boldsymbol{\theta})\boldsymbol{K}\boldsymbol{F}(\boldsymbol{\theta})$. This asymptotic variance has the exact structure as the Marzetta form of the CCRB when $\boldsymbol{K} = \boldsymbol{0}$ and the FIM is full rank. Applying the algebraic identity of Lemma 3.8, then

$$\mathrm{Var}(\sqrt{n}\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} \boldsymbol{U}(\boldsymbol{\theta}_0)\left(\boldsymbol{U}^T(\boldsymbol{\theta}_0)\boldsymbol{D}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{U}^T(\boldsymbol{\theta}_0).$$

It only remains to note

$$\boldsymbol{U}^T(\boldsymbol{\theta}_0)\boldsymbol{D}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0) = \boldsymbol{U}^T(\boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0)$$

since $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{0}$.

Appendix B

Proofs of Convergence Properties of Constrained Scoring

Theorem (Theorem 3.33)**.** If for any iterate $\dot{\boldsymbol{\theta}}^{(k)} \in \Theta_f$ there does not exist an $\alpha^{(k)} > 0$ that satisfies (3.37), then $\dot{\boldsymbol{\theta}}^{(k)}$ is a stationary point.

Proof. Let $\dot{\boldsymbol{\theta}} \in \Theta_f$ and define $\ddot{\boldsymbol{\theta}}(\alpha) = \boldsymbol{\pi}\left[\dot{\boldsymbol{\theta}} + \alpha \mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}})\right]$. By a property of the natural projection of convex sets,

$$\left\|\ddot{\boldsymbol{\theta}}(\alpha) - \dot{\boldsymbol{\theta}}\right\|_{\boldsymbol{I}(\dot{\boldsymbol{\theta}})} \leq \alpha \left\|\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}})\right\|_{\boldsymbol{I}(\dot{\boldsymbol{\theta}})}.$$

Hence it is sufficient to show there exists an $\alpha > 0$ such that

$$\frac{\left(\log p(\boldsymbol{x};\ddot{\boldsymbol{\theta}}(\alpha)) - \log p(\boldsymbol{x};\dot{\boldsymbol{\theta}})\right)}{\alpha} \geq \kappa \boldsymbol{s}^T(\boldsymbol{x};\dot{\boldsymbol{\theta}})\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}).$$

To show this by contradiction, assume not and take the limit as $\alpha \to 0$. Then

$$\begin{aligned} \boldsymbol{s}^T(\boldsymbol{x};\dot{\boldsymbol{\theta}})\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}) &\leq \kappa \boldsymbol{s}^T(\boldsymbol{x};\dot{\boldsymbol{\theta}})\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}) \\ 0 &\leq (\kappa - 1)\boldsymbol{s}^T(\boldsymbol{x};\dot{\boldsymbol{\theta}})\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}). \end{aligned}$$

This inequality implies $\mathrm{CCRB}(\dot{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}) = \boldsymbol{0}$ and $\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}) \in \mathrm{span}(\boldsymbol{F}^T(\dot{\boldsymbol{\theta}}))$ since $\kappa < 1$, so $\dot{\boldsymbol{\theta}}$ satisfies the stationarity condition (3.32). □

Theorem (Theorem 3.34)**.** The sequence $\{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\}$ is a monotone increasing sequence. Furthermore, if $p(\boldsymbol{x};\cdot)$ is bounded above, then $\{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\}$ converges.

Proof. Since $\kappa \geq 0$ and $\|\dot{\boldsymbol{\theta}}^{(k+1)} - \dot{\boldsymbol{\theta}}^{(k)}\|^2_{\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})} \geq 0$, then by the rule in (3.37), the value of the likelihood function can only increase after each iteration. The second

statement is a consequence of the monotone convergence principle, i.e., a bounded monotone sequence converges [38, p. 44, theorem 2-6]. □

Theorem (Theorem 3.35)**.** If the likelihood $p(\boldsymbol{x};\cdot)$ is bounded above, then the sequence

$$\{\log p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k+1)}) - \log p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\}$$

vanishes.

Proof. Since $\{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\}$ converges, then $\frac{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k+1)})}{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})} \to 1$.[1] And since $\log(\cdot)$ is continuous, then $\log \frac{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k+1)})}{p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})} \to 0$. □

Theorem (Theorem 3.36)**.** If the likelihood $p(\boldsymbol{x};\cdot)$ is bounded above, then the sequence

$$\{\|\dot{\boldsymbol{\theta}}^{(k+1)} - \dot{\boldsymbol{\theta}}^{(k)}\|_{\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})}\}$$

vanishes as $k \to \infty$.

Proof. Again, by the rule in (3.37), the sequence $\{\|\dot{\boldsymbol{\theta}}^{(k+1)} - \dot{\boldsymbol{\theta}}^{(k)}\|^2_{\boldsymbol{I}(\dot{\boldsymbol{\theta}}^{(k)})}\}$ is bounded above by the product of a bounded sequence $\{\alpha^{(k)}\}$ and a vanishing sequence $\{\log p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k+1)}) - \log p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})\}$, and clearly each element of the sequence is non-negative. Hence, by the squeezing theorem[2], the sequence vanishes as $k \to \infty$. □

Theorem (Theorem 3.37)**.** If $\Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$ is compact and convex, then limit points of the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$ are also stationary points.

[1] If $a_k \to a$, $b_k \to b$, and $b_k \neq 0$ for any k, then $\frac{a_k}{b_k} \to \frac{a}{b}$ [38, p. 41, theorem 2-4(d)].
[2] If $0 \leq a_k \leq b_k$ and $b_k \to 0$, then $a_k \to 0$ [38, p. 43, theorem 2-5(c)].

Proof. Let $\boldsymbol{\theta}^\star$ be a limit point of the sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$. Then by virtue of Bolzano-Weierstrass [38, p. 54, theorem 2-14], there exists a convergent subsequence $\{\dot{\boldsymbol{\theta}}^{(k_i)}\}$ that converges to $\boldsymbol{\theta}^\star$. Additionally, since $\{\alpha^{(k_i)}\}$ is a bounded sequence, it contains a convergent subsequence $\{\alpha^{(k_{i_j})}\}$ with a limit point we shall denote $\alpha^\star$. It is still true that $\dot{\boldsymbol{\theta}}^{(k_{i_j})} \to \boldsymbol{\theta}^\star$ [38, p.49, theorem 2-10]. Then we can bound the norm-distance between $\boldsymbol{\pi}\left[\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right]$ and $\boldsymbol{\theta}^\star$ using the triangle inequality, e.g.,

$$
\begin{aligned}
&\left\|\boldsymbol{\pi}\left[\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right] - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\quad\leq \left\|\boldsymbol{\pi}\left[\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right] - \boldsymbol{\pi}\left[\dot{\boldsymbol{\theta}}^{(k_{i_j})} + \alpha^{(k_{i_j})} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right]\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\qquad + \left\|\boldsymbol{\pi}\left[\dot{\boldsymbol{\theta}}^{(k_{i_j})} + \alpha^{(k_{i_j})} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right] - \dot{\boldsymbol{\theta}}^{(k_{i_j})}\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} + \left\|\dot{\boldsymbol{\theta}}^{(k_{i_j})} - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\quad\leq \left\|\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \dot{\boldsymbol{\theta}}^{(k_{i_j})} - \alpha^{(k_{i_j})} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\qquad + \left\|\dot{\boldsymbol{\theta}}^{(k_{i_j}+1)} - \dot{\boldsymbol{\theta}}^{(k_{i_j})}\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} + \left\|\dot{\boldsymbol{\theta}}^{(k_{i_j})} - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)}.
\end{aligned}
$$

The second inequality is a result of a property of projections on convex sets for the first term and the definition of our method of scoring with constraints for the second term. Note this second term will vanish by theorem 3.36 as $k_{i_j} \to \infty$. Also, the third term will vanish as $k_{i_j} \to \infty$ since $\boldsymbol{\theta}^\star$ is the limit of the sequence $\{\dot{\boldsymbol{\theta}}^{(k_{i_j})}\}$. The first term is bounded, using the triangle inequality again, as in

$$
\begin{aligned}
&\left\|\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \dot{\boldsymbol{\theta}}^{(k_{i_j})} - \alpha^{(k_{i_j})} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\quad\leq \left\|\dot{\boldsymbol{\theta}}^{(k_{i_j})} - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} + \left\|\alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \alpha^{(k_{i_j})} \mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)}.
\end{aligned}
$$

Again, this first term will vanish as $k_{i_j} \to \infty$. This last term, using the triangle

inequality, satisfies

$$\begin{aligned}
&\left\|\alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \alpha^{(k_{i_j})}\mathrm{CCRB}(\dot{\boldsymbol{\theta}}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k_{i_j})})\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\leq \left\|\alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \alpha^{(k_{i_j})}\mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\quad + \left\|\alpha^{(k_{i_j})}\mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) - \alpha^{(k_{i_j})}\mathrm{CCRB}(\boldsymbol{\theta}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^{(k_{i_j})})\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\leq \left\|\alpha^{(k_{i_j})} - \alpha^\star\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \|\mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \\
&\quad + \left\|\alpha^{(k_{i_j})}\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)} \left\|\mathrm{CCRB}(\boldsymbol{\theta}^{(k_{i_j})})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^{(k_{i_j})}) - \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)}.
\end{aligned}$$

The second inequality used the distributive property of norms [19, p.170, theorem 6.9.2]. By compactness, $\{\|\mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)}\}$ is bounded. So the first term above will vanish as $k_{i_j} \to \infty$ since $\alpha^\star$ is the limit of the sequence $\{\alpha^{(k_{i_j})}\}$. Similarly, $\{\|\alpha^{(k_{i_j})}\|_{\boldsymbol{I}(\boldsymbol{\theta}^\star)}\}$ is a bounded sequence, and so the second term above will vanish as $k_{i_j} \to \infty$ since $\mathrm{CCRB}(\boldsymbol{\theta}^{(k_{i_j})}) \to \mathrm{CCRB}(\boldsymbol{\theta}^\star)$ and $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^{(k_{i_j})}) \to \boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)$ by continuity [38, p.78, corollary 4-2]. Therefore,

$$\boldsymbol{\pi}\left[\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)\right] = \boldsymbol{\theta}^\star$$

and one of the following holds:

(a) $\alpha^\star = 0$,

(b) $\mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star) = \mathbf{0}$, or

(c) the step projection $\boldsymbol{\theta}^\star + \alpha^\star \mathrm{CCRB}(\boldsymbol{\theta}^\star)\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)$ is perpendicular to Θ_f at $\boldsymbol{\theta}^\star$.

This last case (c) is impossible since the step is directed by linear combinations of the vectors of $\boldsymbol{U}(\boldsymbol{\theta}^\star)$ which are tangent to the constraint space at $\boldsymbol{\theta}^\star$. This first case (a) implies stationarity by applying continuity on the step size rule condition (3.37)

and then theorem 3.33. And (b) implies $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}^\star)$ is some linear combination of the columns of $\boldsymbol{F}^T(\boldsymbol{\theta}^\star)$, i.e., $\boldsymbol{\theta}^\star$ satisfies (3.32). Therefore, $\boldsymbol{\theta}^\star$ is a stationary point. □

Theorem (Theorem 3.38)**.** If $\Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$ is compact for all sequences in a closed set of Θ_f and if there is a unique limit point $\boldsymbol{\theta}^\star$ for all such sequences then $\lim_{k\to\infty} \dot{\boldsymbol{\theta}}^{(k)} = \boldsymbol{\theta}^\star$ for every sequence $\{\dot{\boldsymbol{\theta}}^{(k)}\}$. Also, $\boldsymbol{\theta}^\star$ is the maximum of $p(\boldsymbol{x};\cdot)$.

Proof. Let $\dot{\boldsymbol{\theta}}^{(2)}$ be any point in the compact set $\Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$. Since $\{\dot{\boldsymbol{\theta}}^{(k)}\}$ resides in a compact set it has a limit point (Bolzano-Weierstrass [38, p. 52, theorem 2-12]), which must be unique and therefore $\lim_{k\to\infty} \dot{\boldsymbol{\theta}}^{(k)} = \boldsymbol{\theta}^\star$. By theorem 3.34, $p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k+d)}) \geq p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)})$, and by continuity $p(\boldsymbol{x};\dot{\boldsymbol{\theta}}^{(k)}) \to p(\boldsymbol{x};\boldsymbol{\theta}^\star)$. Hence, $p(\boldsymbol{x};\boldsymbol{\theta}^\star) \geq p(\boldsymbol{x};\boldsymbol{\theta}')$ for every $\boldsymbol{\theta}' \in \Theta_{\dot{\boldsymbol{\theta}}^{(1)}}$. □

Appendix C

Proofs of Theorems in Chapter 4

Theorem (Theorem 4.7)**.** The CFIM is singular and the dimension of its null space is lower bounded as

$$\text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) \geq \sum_{i=1}^{K}\sum_{j=1}^{K}(L_i - L_j + 1)_+, \tag{C.1}$$

where $(a)_+ = a$ for $a \geq 0$ and $(a)_+ = 0$ for $a < 0$. This limit quantity is the *nullity lower bound (NLB).*

Proof. This proof is by construction. We shall develop a null subspace of the submatrix $\begin{bmatrix}\boldsymbol{Q}_i & \boldsymbol{Q}_j\end{bmatrix}$ of $\boldsymbol{\mathcal{Q}}$. In particular, consider the submatrix of consisting of the ith source elements and j channel elements corresponding to the mth channel, i.e., $\begin{bmatrix}\boldsymbol{S}^{(i)} & \boldsymbol{H}^{(j)}_{(m)}\end{bmatrix}$. There exists three case to consider: (1) $L_i = L_j$, (2) $L_i > L_j$, and (3) $L_i < L_j$. But first, for use in the proof, define the $M(L_j+1) \times (L_j - L_i + 1)_+$ matrix

$$\boldsymbol{\mathcal{H}}^{(i)}_{(j)} = \begin{bmatrix}\boldsymbol{\mathcal{H}}^{(i)}_{(j)1} \\ \boldsymbol{\mathcal{H}}^{(i)}_{(j)2} \\ \vdots \\ \boldsymbol{\mathcal{H}}^{(i)}_{(j)M}\end{bmatrix} \tag{C.2}$$

where $\boldsymbol{\mathcal{H}}^{(i)}_{(j)m}$ is the $(L_j+1)\times(L_j-L_i+1)_+$ matrix

$$\boldsymbol{\mathcal{H}}^{(i)}_{(j)m}=\begin{bmatrix} h^{(i)}_m(0) & 0 & \cdots & 0\\ h^{(i)}_m(1) & h^{(i)}_m(0) & \cdots & 0\\ \vdots & h^{(i)}_m(1) & \ddots & 0\\ h^{(i)}_m(L_i) & \vdots & \cdots & h^{(i)}_m(0)\\ 0 & h^{(i)}_m(L_i) & \cdots & h^{(i)}_m(1)\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & h^{(i)}_m(L_i)\end{bmatrix}. \tag{C.3}$$

Then $\boldsymbol{\mathcal{H}}^{(i)}_{(j)m}=\boldsymbol{h}^{(i)}_m$ and $\boldsymbol{\mathcal{H}}^{(i)}_{(j)}=\boldsymbol{h}^{(i)}$ from the secondary vector-matrix model in (4.6) if $L_i=L_j$. And also define the $(N+L_j)\times(L_i-L_j+1)_+$ matrix

$$\boldsymbol{\mathcal{S}}^{(i)}_{(j)}=\begin{bmatrix} s^{(i)}(-L_j) & s^{(i)}(-L_j-1) & \cdots & s^{(i)}(-L_i)\\ s^{(i)}(-L_j+1) & s^{(i)}(-L_j) & \cdots & s^{(i)}(-L_i+1)\\ \vdots & \vdots & \ddots & \vdots\\ s^{(i)}(N-1) & s^{(i)}(N-2) & \cdots & s^{(i)}(N-L_i+L_j-1)\end{bmatrix}. \tag{C.4}$$

Then $\boldsymbol{\mathcal{S}}^{(i)}_{(j)}=\boldsymbol{s}^{(i)}$ if $L_i=L_j$. Now consider the cases:

(1) From the dual interpretation of the model in (4.3) and (4.6), it is clear that $\boldsymbol{H}^{(j)}_{(m)}\boldsymbol{s}^{(i)}=\boldsymbol{S}^{(i)}\boldsymbol{h}^{(j)}_m$; hence,

$$\begin{bmatrix}\boldsymbol{I}_M\otimes\boldsymbol{S}^{(i)} & \boldsymbol{H}^{(j)}_M\end{bmatrix}\begin{bmatrix}\boldsymbol{\mathcal{H}}^{(j)}_{(i)}\\ -\boldsymbol{\mathcal{S}}^{(i)}_{(j)}\end{bmatrix}=\boldsymbol{0}.$$

Also, $(L_i-L_j+1)_+=\dim(\begin{bmatrix}\boldsymbol{h}^{(j)}\\ -\boldsymbol{s}^{(i)}\end{bmatrix})=1$ unless $\boldsymbol{h}^{(j)}=\boldsymbol{0}$ and $\boldsymbol{s}^{(i)}=\boldsymbol{0}$, in which case, nullity$(\begin{bmatrix}\boldsymbol{I}_M\otimes\boldsymbol{S}^{(i)} & \boldsymbol{H}^{(j)}_M\end{bmatrix})>(L_i-L_j+1)_+$.

(2) $\boldsymbol{\mathcal{H}}^{(j)}_{(i)m}$ in (C.3) has rank (L_i-L_j+1) unless $\boldsymbol{h}^{(j)}_m=\boldsymbol{0}$, which is impossible by definition. Likewise, $\boldsymbol{\mathcal{S}}^{(i)}_{(j)}$ is full column rank unless $\boldsymbol{s}^{(i)}$ has fewer than (L_i-L_j+1) modes or $N<L_i-2L_j+1$ (see theorem 4.3). Therefore, since

$$\begin{bmatrix}\boldsymbol{I}_M\otimes\boldsymbol{S}^{(i)} & \boldsymbol{H}^{(j)}_M\end{bmatrix}\begin{bmatrix}\boldsymbol{\mathcal{H}}^{(j)}_{(i)}\\ -\boldsymbol{\mathcal{S}}^{(i)}_{(j)}\end{bmatrix}=\boldsymbol{0}$$

then nullity$(\begin{bmatrix}\boldsymbol{I}_M\otimes\boldsymbol{S}^{(i)} & \boldsymbol{H}^{(j)}_M\end{bmatrix})\geq(L_i-L_j+1)_+$.

(3) Finally, nullity($\begin{bmatrix} \boldsymbol{I}_M \otimes \boldsymbol{S}^{(i)} & \boldsymbol{H}_M^{(j)} \end{bmatrix}$) $\geq 0 \geq (L_i - L_j + 1)$.

□

Theorem (CFIM NLB necessary conditions, Theorem 4.13)**.** The M-channel K-source FIR system Fisher information matrix has a nullity of exactly the NLB in (4.13) only if

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + \sum_{j=1}^{K} L_j$,

(c) $p_k \geq L_k + 2$ for $k = 1, \ldots, K$ or $p_k \geq 1$ if $L_k = 0$,

(d) $N \geq K + \sum_{j=1}^{K} L_j$, and

(e) $M > K$.

Proof. If any of these conditions fail, then the null space of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ is greater than the NLB.

(a) If $\boldsymbol{H}(z)$ is reducible or not column-reduced, then by theorem 4.2 $\boldsymbol{H}_M$ is not full column rank. Let $\boldsymbol{v} \in \text{null}(\boldsymbol{H}_M)$ and partition $\boldsymbol{v}$ as

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}^{(1)} \\ \boldsymbol{v}^{(2)} \\ \vdots \\ \boldsymbol{v}^{(K)} \end{bmatrix}$$

where $\boldsymbol{v}^{(k)}$ is a $N + L_k$ length vector. (It is assumed that $\boldsymbol{v}$ is independent of $\boldsymbol{s}$ otherwise $E_{\boldsymbol{\vartheta}}\boldsymbol{y} = \boldsymbol{0}$ and the model itself is not identifiable.) Then

$$\boldsymbol{v}^* = \begin{bmatrix} \boldsymbol{0}^{(1)} \\ \boldsymbol{v}^{(1)} \\ \vdots \\ \boldsymbol{0}^{(K)} \\ \boldsymbol{v}^{(K)} \end{bmatrix} \in \text{null}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}))$$

where $\mathbf{0}^{(k)}$ is a $M(L_k+1)$ length zero vector. Assume $\boldsymbol{v}^*$ is a linear combination of the columns of $\boldsymbol{\mathcal{N}}$ in (4.14). Then for some k the columns of $\begin{bmatrix}\boldsymbol{\mathcal{H}}_{(k)}^{(1)} & \boldsymbol{\mathcal{H}}_{(k)}^{(2)} & \cdots & \boldsymbol{\mathcal{H}}_{(k)}^{(K)}\end{bmatrix}$ must be linearly dependent. (Otherwise without this assumption, then $\boldsymbol{v}^* \notin \text{span}(\boldsymbol{\mathcal{N}})$. This corresponds to the $\mathbf{0}^{(k)}$ subvector in $\boldsymbol{v}^*$.) Let

$$\boldsymbol{u}_{(k)} = \begin{bmatrix}\boldsymbol{u}_{(k)}^{(1)} \\ \boldsymbol{u}_{(k)}^{(2)} \\ \vdots \\ \boldsymbol{u}_{(k)}^{(K)}\end{bmatrix} \in \text{null}(\begin{bmatrix}\boldsymbol{\mathcal{H}}_{(k)}^{(1)} & \boldsymbol{\mathcal{H}}_{(k)}^{(2)} & \cdots & \boldsymbol{\mathcal{H}}_{(k)}^{(K)}\end{bmatrix})$$

with $\boldsymbol{u}_{(k)}^{(i)}$ a $(L_k - L_i + 1)_+$ length vector. Then

$$\boldsymbol{U}_{(k)} = \begin{bmatrix}\boldsymbol{U}_{(k)}^{(1)} \\ \boldsymbol{U}_{(k)}^{(2)} \\ \vdots \\ \boldsymbol{U}_{(k)}^{(K)}\end{bmatrix} \in \text{null}(\boldsymbol{H}_M)$$

where

$$\boldsymbol{U}_{(k)}^{(i)} = \begin{bmatrix}\boldsymbol{u}_{(k)}^{(i)T} & 0 & \cdots & 0 \\ 0 & \boldsymbol{u}_{(k)}^{(i)T} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \boldsymbol{u}_{(k)}^{(i)T}\end{bmatrix}_{N+L_i \times N+L_k}.$$

(If $L_i > L_k$ then $\boldsymbol{U}_{(k)}^{(i)}$ is a null matrix. Also, $\boldsymbol{U}_{(k)}^{(k)} = u_{(k)}\boldsymbol{I}_{N+L_k \times N+L_k}$.) The matrix $\boldsymbol{U}_{(k)}$ can be arranged to create $N+L_k$ linearly independent columns in the null space of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$, where the submatrices $\boldsymbol{U}_{(k)}^{(i)}$ correspond to the rows of $\boldsymbol{\mathcal{N}}$ containing $\begin{bmatrix}-\boldsymbol{\mathcal{S}}_{(i)}^{(1)} & -\boldsymbol{\mathcal{S}}_{(i)}^{(2)} & \cdots & -\boldsymbol{\mathcal{S}}_{(i)}^{(K)}\end{bmatrix}$, which has rank at most $\sum_{j=1}^{K}(L_j - L_i+1)_+ \leq \sum_{j=1}^{K}(L_j+1)$. For the columns of $\boldsymbol{\mathcal{N}}$ to be a basis of the null space of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$, it is needed then that $\sum_{j=1}^{K}(L_j+1) \geq \sum_{j=1}^{K}(L_j-L_k+1)_+ \geq N+L_k \geq N+L_i$,

which implies at most $N = K + \sum_{j=1}^{K} L_j$ contradicting theorem 4.12 unless all the channel orders are zero. In this latter case, then note that

$$\boldsymbol{\mathcal{N}} \cdot \left(\boldsymbol{I}_{K\times K} \otimes \begin{bmatrix} u_{(k)}^{(1)} \\ u_{(k)}^{(2)} \\ \vdots \\ u_{(k)}^{(K)} \end{bmatrix} \right)$$

will not be full rank, and hence neither can $\boldsymbol{\mathcal{N}}$ be so.

(b) If $p_{\text{total}} < K + \sum_{k=1}^{K} L_k$, then there exists $\boldsymbol{v} \in \text{null}(\boldsymbol{S})$ where the matrix $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{S}^{(2)} & \cdots & \boldsymbol{S}^{(K)} \end{bmatrix}$. If $\boldsymbol{v}$ is partitioned as

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}^{(1)} \\ \boldsymbol{v}^{(2)} \\ \vdots \\ \boldsymbol{v}^{(K)} \end{bmatrix}$$

where $\boldsymbol{v}^{(k)}$ is a $L_k + 1$ length vector, then

$$\boldsymbol{v}^* = \begin{bmatrix} \mathbf{1}_{M\times 1} \otimes \boldsymbol{v}^{(1)} \\ \mathbf{0}^{(1)} \\ \vdots \\ \mathbf{1}_{M\times 1} \otimes \boldsymbol{v}^{(K)} \\ \mathbf{0}^{(K)} \end{bmatrix} \in \text{null}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}))$$

with $\mathbf{0}^{(j)}$ is an $N + L_j$ length zero vector. If $\boldsymbol{v}^* \in \text{span}(\boldsymbol{\mathcal{N}})$, then for some k we have $\text{rank}(\begin{bmatrix} \boldsymbol{\mathcal{S}}_{(k)}^{(1)} & \boldsymbol{\mathcal{S}}_{(k)}^{(2)} & \cdots & \boldsymbol{\mathcal{S}}_{(k)}^{(K)} \end{bmatrix}) < \sum_{j=1}^{K} (L_j - L_k + 1)_+$. Then by a construction similar to part (a), $\text{nullity}(\boldsymbol{S}) \geq L_k + 1$ and this contribution to the null space of $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ has rank at least $M(L_k + 1)$, with submatrices that coincide with $\begin{bmatrix} \boldsymbol{\mathcal{H}}_{(k)}^{(1)} & \boldsymbol{\mathcal{H}}_{(k)}^{(2)} & \cdots & \boldsymbol{\mathcal{H}}_{(k)}^{(K)} \end{bmatrix}$, which has rank at most $\sum_{j=1}^{K} (L_k - L_j + 1)_+ \leq \sum_{j=1}^{K} (L_k + 1) = K(L_k + 1) < M(L_k + 1)$ for any L_k since $M > K$. Therefore, $\text{nullity}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})) > \sum_{i=1}^{K} \sum_{j=1}^{K} (L_i - L_j + 1)_+$.

(c) If $p_k < L_k + 1$ then by [31, lemma 1], $\boldsymbol{S}^{(k)}$ and, consequently $\boldsymbol{S}$, has a null space and the argument in (b) applies. So assume $p_k = L_k + 1$. If $N < L_k + 1$ (and $L_k \neq 0$) then $\boldsymbol{S}^{(k)}$ has a null space [76, lemma 1], so assume $N \geq L_k + 1$. From [31, the proof of theorem 1], it is possible to construct a $\boldsymbol{v}$ independent from $\boldsymbol{s}^{(k)}$ such that $\text{span}(\boldsymbol{V}) = \text{span}(\boldsymbol{S}^{(k)})$ for $\boldsymbol{V}$ defined similarly as $\boldsymbol{S}^{(k)}$. So for any $\boldsymbol{h}^{(k)}$ there exists a $\boldsymbol{h}^*$ such that $\boldsymbol{H}_M^{(k)}\boldsymbol{v} = (\boldsymbol{I}_{M\times M} \otimes \boldsymbol{V})\boldsymbol{h}^{(k)} = (\boldsymbol{I}_{M\times M} \otimes \boldsymbol{S}^{(k)})\boldsymbol{h}^*$. Therefore both $\begin{bmatrix} -\boldsymbol{v} \\ \boldsymbol{h}^* \end{bmatrix}$ and $\begin{bmatrix} -\boldsymbol{s}^{(k)} \\ \boldsymbol{h}^{(k)} \end{bmatrix}$ reside in $\text{null}(\begin{bmatrix} \boldsymbol{H}_M^{(k)} & (\boldsymbol{I}_{M\times M} \otimes \boldsymbol{S}^{(k)}) \end{bmatrix})$, which increases the nullity lower bound by at least one.

(d) This is a looser bound and hence required by theorem 4.12.

□

Theorem (CFIM NLB sufficiency conditions, theorem 4.14)**.** The M-channel K-source FIR system FIM has a nullity of exactly the NLB in (4.13) if

(a) $\boldsymbol{H}(z)$ is irreducible and column-reduced,

(b) $p_{\text{total}} \geq K + (K+1)\sum_{j=1}^{K} L_j$,

(c) $p_k \geq L_k + 1 + \sum_{j=1}^{K} L_j$ for $k = 1, \ldots, K$,

(d) $N \geq K + (K+2)\sum_{j=1}^{K} L_j$, and

(e) $M > K$.

This result is conceptually easier to prove from yet another alternative matrix model from the ones in (4.3) or (4.6) using $\boldsymbol{S}(n)$ as defined in (4.9). Then define

the $(L_k + 1 + n) \times (n + 1)$ matrix

$$\boldsymbol{H}_{(m)}^{(k)}(n) = \begin{bmatrix} h_m^{(k)}(0) & & \\ \vdots & \ddots & h_m^{(k)}(0) \\ h_m^{(k)}(L_k) & \ddots & \vdots \\ & & h_m^{(k)}(L_k) \end{bmatrix},$$

which is the impulse response for the ith subchannel of the kth source. Then define $\boldsymbol{H}^{(k)}(n) = \left[\boldsymbol{H}_{(1)}^{(k)}(n), \ldots, \boldsymbol{H}_{(M)}^{(k)}(n)\right]$ and the $(K(n+1) + \sum_{k=1}^{K} L_k) \times M(n+1)$ matrix

$$\boldsymbol{H}(n) = \begin{bmatrix} \boldsymbol{H}^{(1)}(n) \\ \cdots \\ \boldsymbol{H}^{(K)}(n) \end{bmatrix}.$$

The observations for the mth channels (receiver) can be collected into the $(N-n) \times (n+1)$ matrix

$$\boldsymbol{Y}_{(m)}(n) = \begin{bmatrix} y_m(n) & \cdots & y_m(0) \\ \vdots & & \vdots \\ y_m(N-1) & \cdots & y_m(N-1-n) \end{bmatrix}$$

with $\boldsymbol{Y}(n) = \left[\boldsymbol{Y}_{(1)}(n), \ldots, \boldsymbol{Y}_{(M)}(n)\right]$. The noise matrix $\boldsymbol{W}(n)$ is defined similarly. Then the alternative model is

$$\boldsymbol{Y}(n) = \boldsymbol{S}(n)\boldsymbol{H}(n) + \boldsymbol{W}(n). \tag{C.5}$$

Before this theorem is proven, a lemma will be needed. This lemma is a generalization of a result in [52, theorem 3]. The proof was originally shown in [49, Appendix].

Lemma C.1. Assume $\boldsymbol{H}(n)$ be full row rank and $\boldsymbol{h}'^{(k)}$ be any nontrivial $M(L_k+1)$ length vector, and define $\boldsymbol{H}'^{(k)}(n^*)$ to be the $L_k+1+n^* \times M(1+n^*)$ matrix composed from $\boldsymbol{h}'^{(k)}$ as in (4.4) and (4.5). Then the following two statements are equivalent:

(i) $\text{corange}\{\boldsymbol{H}'^{(k)}(n^*)\} \subset \text{corange}\{\boldsymbol{H}^{(1)}(n^*), \ldots, \boldsymbol{H}^{(K)}(n^*)\} = \text{corange}\{\boldsymbol{H}(n^*)\}$.

(ii) $\boldsymbol{h}'^{(k)} \in \text{range}\{\boldsymbol{\mathcal{H}}_{(k)}^{(1)}, \ldots, \boldsymbol{\mathcal{H}}_{(k)}^{(K)}\}$.

We shall prove a more general version of the lemma, which is essentially an extension of the identifiability theorem of [52, theorem 3] from the SIMO to the MIMO scenario. First, define $\boldsymbol{h}^{(j)}(l) = \left[h_{(1)}^{(j)}(l), \ldots, h_{(M)}^{(j)}(l)\right]^T$ and the $Mn \times (n+L_j)$ matrix

$$\boldsymbol{H}^{(j)}(n) = \begin{bmatrix} \boldsymbol{h}^{(j)}(L_j) & \boldsymbol{h}^{(j)}(L_j-1) & \cdots & \boldsymbol{h}^{(j)}(0) & & \\ & \boldsymbol{h}^{(j)}(L_j) & \boldsymbol{h}^{(j)}(L_j-1) & \cdots & \boldsymbol{h}^{(j)}(0) & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \boldsymbol{h}^{(j)}(L_j) & \boldsymbol{h}^{(j)}(L_j-1) & \cdots & \boldsymbol{h}^{(j)}(0) \end{bmatrix}$$

so that $\boldsymbol{H}(N) = \left[\boldsymbol{H}^{(1)}(N), \ldots, \boldsymbol{H}^{(K)}(N)\right]$ is similar to $\boldsymbol{H}_M$ in the original model (4.3), except for the order of the rows. Let L satisfy $\min_j \{L_j\} \leq L \leq \max_j \{L_j\}$. Also define $\boldsymbol{\mathcal{H}}_L$ as the $M(L+1) \times (L - L_j + 1)_+$ matrix

$$\boldsymbol{\mathcal{H}}_L^{(j)} = \begin{bmatrix} \boldsymbol{h}^{(j)} & \boldsymbol{0}_{M\times 1} & \cdots & \boldsymbol{0}_{M(L-L_j-1)\times 1} \\ \boldsymbol{0}_{M\times 1} & \boldsymbol{h}^{(j)} & \ddots & \boldsymbol{0}_{M\times 1} \\ \boldsymbol{0}_{M(L-L_j-1)\times 1} & \boldsymbol{0}_{M(L-L_j-1)\times 1} & & \boldsymbol{h}^{(j)} \end{bmatrix}$$

where $\boldsymbol{h}^{(j)} = \left[\boldsymbol{h}^{(j)}(0)^T, \ldots, \boldsymbol{h}^{(j)}(L_j)^T\right]^T$, an $M(L_j+1) \times 1$ matrix. (The matrix is not to scale, i.e., $\boldsymbol{0}_{M\times 1}$ does not align with $\boldsymbol{h}^{(j)}$.) Note, that $\boldsymbol{\mathcal{H}}_L^{(j)}$ is null for $L < L_j$. Now, we restate the lemma as the following. Instead, we show the following are equivalent:

(1) $\text{Range}\{\boldsymbol{H}'(N)\} \subset \text{Range}\left\{\boldsymbol{H}^{(1)}(N), \ldots, \boldsymbol{H}^{(K)}(N)\right\} = \text{Range}\{\boldsymbol{H}(N)\}$.

(2) $\boldsymbol{h}' \in \text{Range}\left\{\boldsymbol{\mathcal{H}}_L^{(1)}, \ldots, \boldsymbol{\mathcal{H}}_L^{(K)}\right\}$.

Proof. Assume $N > \sum_{j=1}^{K} L_j$ and $\boldsymbol{H}(N-1)$ is full-column rank. Let $\boldsymbol{h}'$ be any $M(L+1) \times 1$ nonzero complex vector, and define $\boldsymbol{H}'(N)$ to be the $MN \times N + L$ channel matrix composed from $\boldsymbol{h}'$. We'll first assume (1), and show (2). Note that

$$\boldsymbol{H}^{(j)}(N) = \begin{bmatrix} \boldsymbol{h}^{(j)}(0) & \boldsymbol{p}^{(j)}(N-1) \\ \boldsymbol{0}_{M(N-1)\times 1} & \boldsymbol{H}^{(j)}(N-1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}^{(j)}(N-1) & \boldsymbol{0}_{M(N-1)\times 1} \\ \boldsymbol{q}^{(j)}(N-1) & \boldsymbol{h}^{(j)}(L_j) \end{bmatrix}$$

where

$$\begin{aligned}\boldsymbol{p}^{(j)}(N-1) &= \left[\boldsymbol{h}^{(j)}(1),\ldots,\boldsymbol{h}^{(j)}(L_j),\mathbf{0}_{M\times N-1}\right], \text{ and}\\ \boldsymbol{q}^{(j)}(N-1) &= \left[\mathbf{0}_{M\times N-1},\boldsymbol{h}^{(j)}(0),\ldots,\boldsymbol{h}^{(j)}(L_j-1)\right].\end{aligned}$$

Note statement (1) implies the first column of $\boldsymbol{H}'(N)$ satisfies

$$\begin{bmatrix}\boldsymbol{h}'(0)\\ \mathbf{0}_{M(N-1)\times 1}\end{bmatrix} = \sum_{j=1}^{K}\begin{bmatrix}\boldsymbol{h}^{(j)}(0) & \boldsymbol{p}^{(j)}(N-1)\\ \mathbf{0}_{M(N-1)\times 1} & \boldsymbol{H}^{(j)}(N-1)\end{bmatrix}\cdot\begin{bmatrix}\alpha_0^{(j)}\\ \boldsymbol{a}_0^{(j)}\end{bmatrix}$$

where $\alpha_0^{(j)}$ is a constant and $\boldsymbol{a}_0^{(j)}$ is an $N-1+L_j\times 1$ vector. Hence, we have the following two linear systems:

(A1) $\boldsymbol{h}'(0) = \sum_{j=1}^{K}\alpha_0^{(j)}\boldsymbol{h}^{(j)}(0) + \boldsymbol{p}^{(j)}(N-1)\boldsymbol{a}_0^{(j)}$

(A2) $\mathbf{0}_{M(N-1)\times 1} = \sum_{j=1}^{K}\boldsymbol{H}^{(j)}(N-1)\boldsymbol{a}_0^{(j)}$

But $\boldsymbol{H}(N-1)$ is full-column rank, thus $\boldsymbol{a}_0^{(j)} = \mathbf{0}_{N+L_j-1\times 1}$ for all j. Therefore,

$$\boldsymbol{h}'(0) = \sum_{j=1}^{K}\alpha_0^{(j)}\boldsymbol{h}^{(j)}(0). \tag{C.6}$$

Likewise, the next column of $\boldsymbol{H}'(N)$ is given by

$$\begin{bmatrix}\boldsymbol{h}'(1)\\ \boldsymbol{h}'(0)\\ \mathbf{0}_{M(N-2)\times 1}\end{bmatrix} = \sum_{j=1}^{K}\begin{bmatrix}\boldsymbol{h}^{(j)}(0) & \boldsymbol{p}^{(j)}(N-1)\\ \mathbf{0}_{M(N-1)\times 1} & \boldsymbol{H}^{(j)}(N-1)\end{bmatrix}\cdot\begin{bmatrix}\alpha_1^{(j)}\\ \boldsymbol{a}_1^{(j)}\end{bmatrix}$$

where, again, $\alpha_1^{(j)}$ is a constant and $\boldsymbol{a}_1^{(j)}$ is an $N+L_j-1\times 1$ vector. Now, we have the following two systems:

(B1) $\boldsymbol{h}'(1) = \sum_{j=1}^{K}\alpha_1^{(j)}\boldsymbol{h}^{(j)}(0) + \boldsymbol{p}^{(j)}(N-1)\boldsymbol{a}_1^{(j)}$

(B2) $\begin{bmatrix}\boldsymbol{h}'(0)\\ \mathbf{0}_{M(N-2)\times 1}\end{bmatrix} = \sum_{j=1}^{K}\boldsymbol{H}^{(j)}(N-1)\boldsymbol{a}_1^{(j)}$

Note (C.6) implies that $\begin{bmatrix} \boldsymbol{h}'(0) \\ \mathbf{0}_{M(N-2)\times 1} \end{bmatrix} = \sum_{j=1}^{K} \alpha_0^{(j)} \cdot \begin{bmatrix} \boldsymbol{h}^{(j)}(0) \\ \mathbf{0}_{M(N-2)\times 1} \end{bmatrix}$, i.e., it is a linear combination of column vectors of $\boldsymbol{H}(N-1)$. Since $\boldsymbol{H}(N-1)$ is full-column rank, then (B2) implies that $\boldsymbol{a}_1^{(j)} = \begin{bmatrix} \alpha_0^{(j)} \\ 0_{N+L_j-2\times 1} \end{bmatrix}$. Thus, evaluating (B1), we have

$$\boldsymbol{h}'(1) = \sum_{j=1}^{K} \alpha_1^{(j)} \boldsymbol{h}^{(j)}(0) + \alpha_0^{(j)} \boldsymbol{h}^{(j)}(1). \tag{C.7}$$

Continuing in this fashion, we arrive at the general expression

$$\boldsymbol{h}'(l) = \sum_{j=1}^{K} \sum_{i=0}^{l} \alpha_{l-i}^{(j)} \boldsymbol{h}^{(j)}(i) \tag{C.8}$$

for $0 \leq l \leq L$. For convenience, we define $\boldsymbol{h}^{(j)}(i)$ to be null if $i < L_j$ or $i < 0$. Now, consider the $(L+2)$nd column of $\boldsymbol{H}'(N)$,

$$\begin{bmatrix} \mathbf{0}_{M\times 1} \\ \boldsymbol{h}'(L) \\ \vdots \\ \boldsymbol{h}'(0) \\ \mathbf{0}_{M(N-L-2)\times 1} \end{bmatrix} = \sum_{j=1}^{K} \begin{bmatrix} \boldsymbol{h}^{(j)}(0) & \boldsymbol{p}^{(j)}(N-1) \\ \mathbf{0}_{M(N-1)\times 1} & \boldsymbol{H}^{(j)}(N-1) \end{bmatrix} \cdot \begin{bmatrix} \alpha_{L+1}^{(j)} \\ \boldsymbol{a}_{L+1}^{(j)} \end{bmatrix}$$

It follows that

(C1) $\mathbf{0}_{M\times 1} = \sum_{j=1}^{K} \alpha_{L+1}^{(j)} \boldsymbol{h}^{(j)}(0) + \boldsymbol{p}^{(j)}(N-1) \boldsymbol{a}_{L+1}^{(j)}$

(C2) $\begin{bmatrix} \boldsymbol{h}'(L) \\ \vdots \\ \boldsymbol{h}'(0) \\ \mathbf{0}_{M(N-L-2)\times 1} \end{bmatrix} = \sum_{j=1}^{K} \boldsymbol{H}^{(j)}(N) \boldsymbol{a}_{L+1}^{(j)}$

Since $\boldsymbol{H}(N-1)$ is full-column rank, equation (C.8) and (C2) imply that $\boldsymbol{a}_{L+1}^{(j)} = \left[\alpha_L^{(j)}, \ldots, \alpha_0^{(j)}, \mathbf{0}_{1\times M(N-L-2)}\right]^T$. Applied to (C1) we have

$$\mathbf{0}_{M\times 1} = \sum_{j=1}^{K} \alpha_{L+1}^{(j)} \boldsymbol{h}^{(j)}(0) + \alpha_L^{(j)} \boldsymbol{h}^{(j)}(1) + \cdots + \alpha_{L-L_j+1}^{(j)} \boldsymbol{h}^{(j)}(L_j)$$

Continuing in this fashion, we see that for $L+1 \leq l \leq N-1$,

$$\mathbf{0}_{M\times 1} = \sum_{j=1}^{K} \sum_{i=0}^{L_j} \alpha_{l-i}^{(j)} \boldsymbol{h}^{(j)}(i) \tag{C.9}$$

In particular, (C.8) and (C.9) show that the Nth column of $\boldsymbol{H}'(N)$ can be written as

$$\begin{bmatrix} \mathbf{0}_{M(N-L-1)\times 1} \\ \boldsymbol{h}'(L) \\ \vdots \\ \boldsymbol{h}'(0) \end{bmatrix} = \sum_{j=1}^{K} \boldsymbol{H}^{(j)}(N) \cdot \begin{bmatrix} \alpha^{(j)}(N-1) \\ \vdots \\ \alpha_0^{(j)} \\ \mathbf{0}_{L_j\times 1} \end{bmatrix} \tag{C.10}$$

Similarly, statement (1) implies the last column of $\boldsymbol{H}'(N)$ satisfies

$$\begin{bmatrix} \mathbf{0}_{M(N-1)\times 1} \\ \boldsymbol{h}'(L) \end{bmatrix} = \sum_{j=1}^{K} \begin{bmatrix} \boldsymbol{H}^{(j)}(N-1) & \mathbf{0}_{M(N-1)\times 1} \\ \boldsymbol{q}^{(j)}(N-1) & \boldsymbol{h}^{(j)}(L_j) \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{y}_L^{(j)} \\ \beta_L^{(j)} \end{bmatrix}$$

where $\boldsymbol{y}_L^{(j)}$ is an $N+L_j+1\times 1$ vector and $\beta_L^{(j)}$ a constant. Thus we have

(a1) $\mathbf{0}_{M(N-1)\times 1} = \sum_{j=1}^{K} \boldsymbol{H}^{(j)}(N-1)\boldsymbol{y}_L^{(j)}$

(a2) $\boldsymbol{h}'(L) = \sum_{j=1}^{K} \boldsymbol{q}^{(j)}(N-1)\boldsymbol{y}_L^{(j)} + \beta_L^{(j)}\boldsymbol{h}^{(j)}(L_j)$

Here, since $\boldsymbol{H}(N-1)$ is full rank, then (a1) implies that $\boldsymbol{y}_L^{(j)} = \mathbf{0}_{N+L_j-1\times 1}$, hence

$$\boldsymbol{h}'(L) = \sum_{j=1}^{K} \beta_L^{(j)} \boldsymbol{h}^{(j)}(L_j)$$

Proceeding as before, it is clear that

$$\boldsymbol{h}'(L-l) = \sum_{j=1}^{K} \sum_{i=0}^{l} \beta_{L-l}^{(j)} \boldsymbol{h}^{(j)}(L_j - l + i)$$

for $0 \leq l \leq L$. Thus, the Nth column of $\boldsymbol{H}'(N)$ can also be expressed as

$$\begin{bmatrix} \mathbf{0}_{M(N-L-1)\times 1} \\ \boldsymbol{h}'(L) \\ \vdots \\ \boldsymbol{h}'(0) \end{bmatrix} = \sum_{j=1}^{K} \boldsymbol{H}^{(j)}(N) \cdot \begin{bmatrix} \mathbf{0}_{N+L_j-L-1\times 1} \\ \beta_L^{(j)} \\ \vdots \\ \beta_0^{(j)} \end{bmatrix} \tag{C.11}$$

But $\boldsymbol{H}(N)$ is full-column rank, hence (C.10) and (C.11) imply that

$$\begin{bmatrix} \alpha^{(j)}(N-1) \\ \vdots \\ \alpha_0^{(j)} \\ \mathbf{0}_{L_j \times 1} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N+L_j-L-1 \times 1} \\ \beta_L^{(j)} \\ \vdots \\ \beta_0^{(j)} \end{bmatrix}$$

Thus, $\alpha_l^{(j)} = 0$ for all $l \geq L+1$. Now, consider the following three cases:

(i) $L = L_j$, then $\alpha_0^{(j)} = \beta_L^{(j)}$ and $\alpha_l^{(j)}, \beta_{L-l}^{(j)} = 0$ for $1 \leq l \leq L$.

(ii) $L < L_j$, then $\alpha_l^{(j)}, \beta_{L-l}^{(j)} = 0$ for $0 \leq l \leq L$.

(iii) $L > L_j$, then $\alpha_l^{(j)} = \beta_{L_j+l}^{(j)}$ for $0 \leq l \leq L-L_j$ and $\alpha_l^{(j)}, \beta_{L-l}^{(j)} = 0$ for $L-L_j+1 \leq l \leq L$.

Define an $(L-L_j+1)^+ \times 1$ vector $\boldsymbol{\alpha}^{(j)} = \left[\alpha_0^{(j)}, \ldots, \alpha_{L-L_j}^{(j)}\right]^T$ for each j. If $L < L_j$, let $\boldsymbol{\alpha}$ be null. Then,

$$\boldsymbol{h}' = \sum_{j=1}^{K} \boldsymbol{\mathcal{H}}_L^{(j)} \boldsymbol{\alpha}^{(j)}$$

or $\boldsymbol{h}' \in \text{Range}\left\{\boldsymbol{\mathcal{H}}_L^{(1)}, \ldots, \boldsymbol{\mathcal{H}}_L^{(K)}\right\}$. This proves statement (2).

Now, conversely assume statement (2). Then $\boldsymbol{h}' = \sum_{j=1}^{K} \boldsymbol{\mathcal{H}}_L^{(j)} \boldsymbol{\gamma}^{(j)}$ where $\boldsymbol{\gamma}^{(j)} = \left[\gamma_1^{(j)}, \ldots, \gamma_{(L-L_j+1)^+}^{(j)}\right]^T$ is an $(L-L_j+1)^+ \times 1$ vector. Define

$$\boldsymbol{\Gamma}^{(j)}(N) = \begin{bmatrix} \gamma_1^{(j)} & \gamma_2^{(j)} & \cdots & \gamma_{(L-L_j+1)^+}^{(j)} & & \\ & \gamma_1^{(j)} & \gamma_2^{(j)} & \cdots & \gamma_{(L-L_j+1)^+}^{(j)} & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \gamma_1^{(j)} & \gamma_2^{(j)} & \cdots & \gamma_{(L-L_j+1)^+}^{(j)} \end{bmatrix}_{N+L_j \times N+L}$$

Note, if $L < L_j$, then $\boldsymbol{\Gamma}^{(j)}(N) = \mathbf{0}_{N+L_j \times N+L}$. Then, it is simple to verify that

$$\boldsymbol{H}'(N) = \sum_{j=1}^{K} \boldsymbol{H}^{(j)}(N) \boldsymbol{\Gamma}^{(j)}(N)$$

which proves statement (1). □

Now having this lemma at hand, theorem 4.14 can be proven.

Proof. Let $n^* = \sum_{j=1}^{K} L_j$. If (a)-(d) are satisfied, then we have column-rank$(\boldsymbol{S}(n^*)) =$ row-rank$(\boldsymbol{H}(n^*) = K + (K+1)n^*$. Let $\boldsymbol{v} \in \text{null}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}))$ where $\boldsymbol{v}$ is partitioned as

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{h}'^{(1)} \\ \boldsymbol{s}'^{(1)} \\ \vdots \\ \boldsymbol{h}'^{(K)} \\ \boldsymbol{s}'^{(K)} \end{bmatrix}.$$

Then $\sum_{k=1}^{K} (\boldsymbol{I}_{M\times M} \otimes \boldsymbol{S}^{(k)}) \boldsymbol{h}'^{(k)} + \boldsymbol{H}_M^{(k)} \boldsymbol{s}'^{(k)} = \boldsymbol{0}$, or in the alternative model in (C.5), we have $\sum_{k=1}^{K} \boldsymbol{S}^{(k)}(n^*) \boldsymbol{H}'^{(k)}(n^*) + \sum_{k=1}^{K} \boldsymbol{S}'^{(k)}(n^*) \boldsymbol{H}^{(k)}(n^*) = \boldsymbol{0}$, or

$$\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right] \begin{bmatrix} \boldsymbol{H}'(n^*) \\ \boldsymbol{H}(n^*) \end{bmatrix} = \boldsymbol{0}.$$

Therefore nullity$(\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right]) \geq$ rank$(\begin{bmatrix} \boldsymbol{H}'(n^*) \\ \boldsymbol{H}(n^*) \end{bmatrix})$. Since

$$\begin{aligned} \text{nullity}(\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right]) &= \text{columns}(\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right]) - \text{rank}(\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right]) \\ &\leq 2K + 2(K+1)n^* - \text{rank}(\boldsymbol{S}(n^*)) \\ &= K + (K+1)n^* \end{aligned}$$

and

$$\text{rank}(\begin{bmatrix} \boldsymbol{H}'(n^*) \\ \boldsymbol{H}(n^*) \end{bmatrix}) \geq \text{rank}(\boldsymbol{H}(n^*)) = K + (K+1)n^*,$$

then nullity$(\left[\boldsymbol{S}(n^*) \boldsymbol{S}'(n^*)\right]) =$ rank$(\begin{bmatrix} \boldsymbol{H}'(n^*) \\ \boldsymbol{H}(n^*) \end{bmatrix}) = K + (K+1)n^*$. Since rank$(\boldsymbol{H}(n^*))$ $= K + (K+1)n^*$ then there exist some matrix $\boldsymbol{T}$ such that $\boldsymbol{H}'(n^*) = \boldsymbol{T}\boldsymbol{H}(n^*)$. Thus, by the lemma, for each k then $\boldsymbol{h}'^{(k)} = \sum_{j=1}^{K} \boldsymbol{\mathcal{H}}_{(k)}^{(j)} \boldsymbol{\gamma}^{(k,j)}$ for some $(L_k - L_j + 1)_+$ length vector. In the alternative model of (C.5), $\boldsymbol{H}'^{(k)}(n^*) = \sum_{j=1}^{K} \boldsymbol{\Gamma}^{(k,j)} \boldsymbol{H}^{(j)}(n^*)$

where

$$\boldsymbol{\Gamma}^{(k,j)} = \begin{bmatrix} \gamma^{(k,j)}_{(L_k-L_j+1)_+} & & & \\ \vdots & \ddots & & \\ \gamma^{(k,j)}_2 & & \gamma^{(k,j)}_{(L_k-L_j+1)_+} & \\ \gamma^{(k,j)}_1 & & \vdots & \gamma^{(k,j)}_{(L_k-L_j+1)_+} \\ & \ddots & \gamma^{(k,j)}_2 & \vdots \\ & & \gamma^{(k,j)}_1 & \gamma^{(k,j)}_2 \\ & & 0 & \gamma^{(k,j)}_1 \end{bmatrix}_{L_k+1+n^* \times L_j+1+n^*}$$

and this defines $\boldsymbol{T}$ as

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{\Gamma}^{(1,1)} & \boldsymbol{\Gamma}^{(1,2)} & \dots & \boldsymbol{\Gamma}^{(1,K)} \\ \boldsymbol{\Gamma}^{(2,1)} & \boldsymbol{\Gamma}^{(2,2)} & \dots & \boldsymbol{\Gamma}^{(2,K)} \\ \vdots & \vdots & \dots & \vdots \\ \boldsymbol{\Gamma}^{(K,1)} & \boldsymbol{\Gamma}^{(K,2)} & \dots & \boldsymbol{\Gamma}^{(K,K)} \end{bmatrix}.$$

Then since

$$\left(\boldsymbol{S}(n^*)\boldsymbol{T} + \boldsymbol{S}^{'}(n^*)\right)\boldsymbol{H}(n^*) = \begin{bmatrix}\boldsymbol{S}(n^*) & \boldsymbol{S}^{'}(n^*)\end{bmatrix}\begin{bmatrix}\boldsymbol{T}\boldsymbol{H}(n^*) \\ \boldsymbol{H}(n^*)\end{bmatrix} = \boldsymbol{0}$$

and $\boldsymbol{H}(n^*)$ is full row rank, it can be seen that $\boldsymbol{S}^{'}(n^*) = -\boldsymbol{S}(n^*)\boldsymbol{T}$ or $\boldsymbol{S}^{'(k)}(n^*) = \sum_{j=1}^{K} -\boldsymbol{S}^{(j)}(n^*)\boldsymbol{\Gamma}^{(j,k)}$, which implies in the vector-matrix model in (4.3) that $\boldsymbol{s}^{'(k)} = \sum_{j=1}^{K} -\boldsymbol{\mathcal{S}}^{(j)}_{(k)}\boldsymbol{\gamma}^{(j,k)}$. Let

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}^{(1,1)} \\ \boldsymbol{\gamma}^{(1,2)} \\ \boldsymbol{\gamma}^{(2,2)} \\ \boldsymbol{\gamma}^{(2,1)} \\ \vdots \\ \boldsymbol{\gamma}^{(1,K)} \\ \vdots \\ \boldsymbol{\gamma}^{(K,K)} \\ \vdots \\ \boldsymbol{\gamma}^{(K,1)} \end{bmatrix}.$$

Then $\boldsymbol{v} = \boldsymbol{\mathcal{N}}\boldsymbol{\gamma}$ and $\text{span}(\boldsymbol{\mathcal{N}}) \equiv \text{null}(\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}))$. □

Bibliography

[1] K. Abed-Meraim, Y. Hua, "Strict identifiability of multichannel FIR systems: further results and developments," *Proceedings of the International Conference on Telecommunications*, Melbourne, Australia, pp. 1029-1032, April 1997.

[2] John Aitchison, "Large Sample Restricted Parametric Tests," *Journal of the Royal Statistical Society, Series B (Methodological)*, Volume 24, Number 1, pp. 234-250, 1962.

[3] J. Aitchison, S.D. Silvey, "Maximum likelihood estimation of parameters subject to restraints," *Ann. Math. Statist.*, Volume 29, pp. 813-828, 1958.

[4] J. Aitchison, S.D. Silvey, "Maximum-likelihood estimation procedures and associated tests of significance," *Journal of the Royal Statistical Society, Series B (Methodological)*, Volume 22, No. 1, pp. 154-171, 1960.

[5] Joshua N. Ash, "On singular estimation problems in sensor localization systems," Ph.D. Thesis, Ohio State University, 2007.

[6] Joshua N. Ash, Randolph L. Moses, "On the relative error and absolute positioning errors in self-localization systems," *IEEE Transactions on Signal Processing*, Volume 56, Number 11, pp. 5668-5679, November 2008.

[7] W.J. Bangs, *Array Processing with Generalized Beamformers*, Ph.D. Thesis, Yale University, New Haven, CT, 1971.

[8] E. W. Barankin, "Locally best unbiased estimates," *Ann. Math. Statist.*, Volume 20, Number 4, pp. 477-501, 1949.

[9] Zvika Ben-Haim, Yonina Eldar, "On the constrained Cramér-Rao bound with a singular Fisher information matrix," *IEEE Signal Processing Letters*, Volume 16, Number 6, pp. 453-456, June 2009.

[10] Dimitri Bertsekas, *Nonlinear Programming: 2nd Edition*, Athena Scientific, 1999.

[11] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Sankhya*, Volume 8, pp. 1-14, 1946.

[12] A.A. Borovkov, A. Moullagaliev, *Mathematical Statistics*. Boca-Raton, FL: CRC, 1998.

[13] Roger Bowden, "The theory of parametric identification," *Econometrica*, Volume 41, Number 6, pp. 1069-1074, November 1973.

[14] George Casella, Roger L. Berger, *Statistical Inference, 2nd Edition*, Duxbury Press, 2002.

[15] F. Cayre, C. Fontaine, T. Furon, "Watermarking security: theory and practice," *IEEE Transactions on Signal Processing*, Volume 53, Number 10, pp. 3976-3987, October 2005.

[16] D. G. Chapman, H. Robbins, "Minimum variance estimation without regularity assumptions," *Ann. Math. Stat.*, Volume 22, pp. 581-586, 1951.

[17] Harald Cramér, *Mathematical Methods of Statistics*, Princeton NJ: Princeton University Press, 1946.

[18] Martin Crowder, "On constrained maximum likelihood estimation with non-i.i.d. observations," *Ann. Inst. Statist. Math.*, Volume 36, Part A, pp.239-249, 1984.

[19] J.N. Franklin, *Matrix Theory*. New York: Dover, 1993.

[20] James E. Gentle, *Matrix Algebra: Theory, Computations, and Applications in Statistics*. New York: Springer, 2007.

[21] P.E. Gill, W. Murray, and M.H. Wright, *Practical Optimization*. New York: Academic, 1981.

[22] A.A. Goldstein, "Convex programming in Hilbert space," *Bulletin of the American Mathematics Society*, Volume 70, Number 5, pp. 709-710, 1964.

[23] John D. Gorman, Alfred O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Transactions on Information Theory*, Volume 26, Number 6, pp. 1285-1301, November 1990.

[24] A. Gorokhov, P. Loubaton, "Subspace based techniques for blind separation of convolutive mixtures with temporally correlated sources," *IEEE Transactions on Circuits and Systems*, Volume 44, pp. 813-820, September 1997.

[25] R.T. Haftka, Z. Gürdal, *Elements of Structural Optimization*. Norwell, MA: Kluwer Academic, 1992.

[26] J. M. Hammersley, "On estimating restricted parameters," *Journal of the Royal Statistical Society, Series B*, Volume 12, Number 2, pp. 192-240, 1950.

[27] Harrie Hendriks, "A Cramér-Rao type lower bound for estimators with values in a manifold," *Journal of Multivariate Analysis*, Volume 38, pp. 245-261, 1991.

[28] A.O. Hero, R. Piramuthu, J.A. Fessler, S.R. Titus, "Minimax emission computed tomography using high-resolution anatomical side information and B-spline models," *IEEE Transactions on Information Theory*, Volume 45, Number 3, pp. 920-938, April 1999.

[29] Bertrand Hochwald, Arye Nehorai, "On identifiability and information-regularity in parameterized normal distributions," *Circuits Systems Signal Processing*, Volume 16, Number 1, pp. 83-89, 1997.

[30] Yingbo Hua, "Fast maximum-likelihood for blind identification of multiple FIR channels," *IEEE Transactions on Signal Processing*, Volume 44, Number 3, pp. 661-672, March 1996.

[31] Y. Hua, M. Wax, "Strict identifiability of multiple FIR channels driven by an unknown arbitrary sequence," *IEEE Transactions on Signal Processing*, Volume 44, pp. 756-759, March 1996.

[32] Aditya K. Jagannatham, Bhaskar D. Rao, "Cramér-Rao lower bound for constrained complex parameters," *IEEE Signal Processing Letters*, Volume 11, Number 11, pp. 875-878, November 2004.

[33] Mortaza Jamshidian, "On algorithms for restricted maximum likelihood estimation," *Computational Statistics & Data Analysis*, Volume 45, pp. 137-157, 2004.

[34] Jürgen Jost, *Riemannian Geometry and Geometric Analysis*, Third Edition, Springer-Verlag, 2002.

[35] T. Kailath, *Linear Systems*. Englewood Cliffs, NH: Prentice-Hall, 1980.

[36] Steven M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[37] C.G. Khatri, "A note on a MANOVA model applied to problems in growth curve," *Annals of the Institute of Statistical Mathematics*, Volume 18, pp. 75-86, 1966.

[38] J.R. Kirkwood, *An Introduction to Analysis*. Boston, MA: PWS Kent, 1995.

[39] Richard J. Kozick, Brian M. Sadler, "Maximum likelihood array processing: The semi-blind case," *Proceedings of the 2nd IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp.70-73, 1999.

[40] Amir Leshem, Alle-Jan van der Veen, "Direction-of-arrival estimation for constant modulus signals," *IEEE Transactions on Signal Processing*, Volume 47, Number 11, pp. 3125-2319, November 1999.

[41] Jian Li, R.T. Compton, Jr., "Maximum likelihood angle estimation for signals with known waveforms," *IEEE Transactions on Signal Processing*, Volume 41, Number 9, pp. 2850-2863, September 1993.

[42] Z. Liu, G.B. Giannakis, S. Barbarossa, A. Scaglione, "Transmit antennae space-time block coding for generalized OFDM in the presence of unknown multipath," *IEEE Journal on Selected Areas in Communications*, Volume 19, Number 7, pp. 1352-1364, July 2001.

[43] H. Liu, G. Xu, L. Tong, "A deterministic approach to blind channel identification of multi-channel FIR systems," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Adelaide, Australia, Volume 4, pp. 581-584, April 1994.

[44] P. Loubaton, E. Moulines, "On blind multiuser forward link channel estimation by the subspace method: Identifiability results," *IEEE Transactions on Signal Processing*, Volume 48, pp. 2366-2376, August 2000.

[45] D.G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1973.

[46] J.H. Manton, "An improved least squares blind channel identification algorithm for linear and affinely precoded communication systems," *IEEE Signal Processing Letters*, Volume 9, Number 9, pp. 282-285, September 2002.

[47] Thomas L. Marzetta, "A simple derivation of the constrained multiple parameter Cramér-Rao bound," *IEEE Transactions on Signal Processing*, Volume 41, Number 6, pp. 2247-2249, June 1993.

[48] Terrence J. Moore, Brian M. Sadler, Richard J. Kozick, "Regularity and strict identifiability in MIMO systems," *IEEE Transactions on Signal Processing*, Volume 50, Number 8, pp. 1831-1842, August 2002.

[49] Terrence J. Moore, Brian M. Sadler, "Sufficient conditions for regularity and strict identifiability in MIMO systems," *IEEE Transactions on Signal Processing*, Volume 52, Number 9, pp. 2650-2655, September 2004.

[50] Terrence J. Moore, Richard J. Kozick, Brian M. Sadler, "The constrained Cramér-Rao bound from the perspective of fitting a model," *IEEE Signal Processing Letters*, Volume 14, Number 8, pp. 564-567, August 2007.

[51] Terrence J. Moore, Brian M. Sadler, Richard J. Kozick, "Maximum-likelihood estimation, the Cramér-Rao bound, and the method of scoring with parameter constraints," *IEEE Transactions on Signal Processing*, Volume 56, Number 3, pp. 895-908, March 2008.

[52] E. Moulines, P. Duhamel, J. Cardoso, S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Transactions on Signal Processing*, Volume 43, pp. 516-525, February 1995.

[53] J.M. Oller, J.M. Corcuera, "Intrinsic analysis of statistical estimation," *The Annals of Statistics*, Volume 23, Number 5, pp. 1562-1581, October 1995.

[54] M.R. Osborne, "Fisher's method of scoring," *Int. Stat. Rev.*, Volume 60, pp. 99-117, 1992.

[55] M.R. Osborne, "Scoring with constraints," *ANZIAM Journal*, Volume 42, Number 1, pp. 9-25, July 2000.

[56] Calyampudi Radakrishna Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, Volume 37, pp. 81-89, 1945.

[57] Alvin C. Rencher, *Linear Models In Statistics*, John Wiley & Sons, New York, 2000.

[58] Thomas J. Rothenberg, "Identification in parametric models," *Econometrica*, Volume 39, Number 3, pp. 577-591, May 1971.

[59] Brian M. Sadler, Richard J. Kozick, Terrence Moore, "Bounds on bearing and symbol estimation with side information," *IEEE Transactions on Signal Processing*, Volume 49, Number 4, pp. 822-834, April 2001.

[60] L.L. Scharf, L.T. McWhorter, "Geometry of the Cramér-Rao bound," *Conference Proceedings, IEEE Sixth SP Workshop on Statistical Signal and Array Processing*, pp. 5-8, October 1992.

[61] George A.F. Seber, Alan J. Lee, *Linear Regression Analysis: 2nd Edition*, Wiley-Interscience, 2003.

[62] Jun Shao, *Mathematical Statistics*. New York, NY: Springer-Verlag, 2003.

[63] S.D. Silvey, "The Lagrange multiplier test," *Ann. Math. Statist.*, Volume 30, pp. 389-407, 1959.

[64] D.R. Smart, *Fixed Point Theorems*, Cambridge University Press, 1980.

[65] M. Spivak, *Calculus on Manifolds*. Reading, MA: Addison-Wesley, 1965.

[66] Petre Stoica, Thomas L. Marzetta, "Parameter estimation problems with singular information matrices," *IEEE Transactions on Signal Processing*, Volume 49, Number 1, pp. 87-90, January 2001.

[67] Petre Stoica, Arye Nehorai, "MUSIC, maximum likelihood, and Cramér-Rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume 37, Number 5, pp. 720-741, May 1989.

[68] Petre Stoica, Boon Chong Ng, "On the Cramér-Rao bound under parametric constraints," *IEEE Signal Processing Letters*, Volume 5, Number 7, pp. 177-179, July 1998.

[69] Petre Stoica, Torsten Söderström, "On non-singular information matrices and local identifiability," *International Journal on Control*, Volume 36, Number 2, pp. 323-329, 1982.

[70] A. van den Bos, "A Cramér-Rao lower bound for complex parameters," *IEEE Transactions on Signal Processing*, Volume 42, Number 10, p. 2859, October 1994.

[71] Alle-Jan van der Veen, "Asymptotic properties of the algebraic constant modulus algorithm," *IEEE Transactions on Signal Processing*, Volume 49, Number 8, pp. 1796-1807, August 2001.

[72] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley and Sons, Inc. 1968.

[73] L.J. Waldorp, H.M. Huizenga, R.P.P.P. Grasman, "The Wald test and Cramér-Rao bound for misspecified models in electromagnetic source analysis," *IEEE Transactions on Signal Processing*, Volume 53, Number 9, pp. 3427-3435, September 2005.

[74] João Xavier, Victor Barroso, "Intrinsic distance lower bound for unbiased estimators on Riemannian manifolds," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, Volume 2, pp. 1141-1144, May 2002.

[75] João Xavier, Victor Barroso, "Intrinsic variance lower bound (IVLB): An extension of the Cramér-Rao bound to Riemannian manifolds," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, Volume 5, pp. 1033-1036, March 2005.

[76] Guanghan Xu, Hui Liu, Lang Tong, Thomas Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, Volume 43, Number 12, pp. 2982-2993, December 1995.

[77] Yingwei Yao, Georgios Giannakis, "On regularity and identifiability of blind source separation under constant-modulus constraints," *IEEE Transactions on Signal Processing*, Volume 53, Number 4, pp. 1272-1281, April 2005.

[78] J. Ziv, M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Transactions on Information Theory*, Volume IT-15, pp. 386-391, May 1969.